Applied Bioinformatics for ncRNA Characterization

Case Studies Combining Next Generation Sequencing & Genomics

Dissertation
zur Erlangung des Doktorgrades
der Naturwissenschaften
(Dr. rer. nat.)

dem Fachbereich Pharmazie der
Philipps-Universität Marburg
vorgelegt von

Ms. Sc. **Clemens Thölken**
aus Göttingen

Marburg (Lahn) 2018

Erstgutachter: **Prof. Dr. Roland K. Hartmann**

Zweitgutachter: **Dr. Marcus Lechner**

Eingereicht am 5.11.2018

Tag der mündlichen Prüfung am 19.12.2018

Hochschulkennziffer: 1180

Erklärung

Ich versichere, dass ich meine Dissertation

„Applied Bioinformatics for ncRNA Characterization – Case Studies Combining Next Generation Sequencing & Genomics“

selbständig ohne unerlaubte Hilfe angefertigt und mich dabei keiner anderen als der von mir ausdrücklich bezeichneten Quellen bedient habe. Alle vollständig oder sinngemäß übernommenen Zitate sind als solche gekennzeichnet.

Die Dissertation wurde in der jetzigen oder einer ähnlichen Form noch bei keiner anderen Hochschule eingereicht und hat noch keinen sonstigen Prüfungszwecken gedient.

Marburg, den 5. November 2018

.....
(Clemens Thölken)

Summary

Non-coding RNAs (ncRNAs) present a diverse class of functional molecules inherent in virtually all forms of cellular life. Besides the canonical protein-encoding mRNAs the role of these abundant transcripts has been overlooked for decades. Defined by their highly conserved structure ncRNAs are resistant to degradation and perform various regulatory functions. Despite the poor sequence conservation, comparative genomics can be employed to identify homologous ncRNAs based on their structure in related species. Through the availability of next generation sequencing techniques, a rich corpus of datasets is available which grants a detailed look into cellular processes. The combination of genomic and transcriptomic data allows for a detailed understanding of molecular mechanism as well as characterization of individual gene functions and their evolution. However, analytical processing of modern high-throughput data is only made viable through optimized bioinformatic algorithms and reproducible automation pipelines.

This thesis consists of four major parts highlighting the diverse roles of ncRNAs concerning the transcription process viewed from different vantage points. The first part concerns an unusually long untranslated region in *Rhodobacter* which harbors a ncRNA that regulates the expression of the downstream division cell wall cluster. Second, the degradation of 6S RNA in *Bacillus subtilis* is experimentally reconstructed to shed light on this final part of the RNA life cycle. This ncRNA is ubiquitous among bacteria and known to be a global transcription regulator itself. Next, the focus moves to the eukaryotic system and RNase P, an ancient ribozyme that is involved in tRNA maturation. Due to differences in composition with an optional RNA and multiple protein subunits, its phylogenetic distribution and deviant characteristics throughout the eukaryotic lineage are examined in order to trace its evolution. Finally, a diverse subgroup of non-translated RNAs are circRNAs which recently received increased attention due to their abundance in neural tissue. Resulting from post-transcriptional back-splicing events circRNAs compete with their host gene for expression. In a zoological study of social insects circRNA were for the first time identified in honeybees. The goal was to find task-related differences in circRNA expression between nurse bees and foragers and thus pinpoint potential functions of these elusive ncRNAs.

The combination of genomic methods and transcriptomic data makes in-depth functional analysis of ncRNAs possible and enables us to understand the molecular mechanisms on multiple levels. Through structural predictions a riboswitch like transcriptional control of UpsM was revealed that is unique to Rhodobacteraceae. Transcriptomic analysis exposed that 6S RNA is primarily processed by RNase J1 for maturation and degraded at internal loops by RNase Y. Evolutionary comparison of organellar RNase P revealed that the RNA subunit is potentially less conserved than thought while organellar protein-only variants are widespread potentially due to horizontal gene transfer. In the case of circRNA, an entire group of ncRNAs was characterized in the social model organism of honeybees and evidence of at least one gene where circRNA levels are significantly reduced during nurse-to-forager transition could be shown. Moreover, an unexpected link between elevated DNA methylation and RNA circularization was discovered. The bioinformatic findings in all of these cases provide a foundation for further experimental research and illustrate how scientific endeavors cannot be automated completely but require rigorous investigation with customized tools.

Zusammenfassung

Nicht-kodierende RNAs (ncRNAs) sind eine verbreitete Klasse von funktionellen Molekülen und praktisch in allen Formen zellulären Lebens anzutreffen. Neben kanonisch protein-enkodierenden mRNAs wurde die Rolle dieser stark vertretenen Transkripte allerdings lange übersehen. Primär definiert durch ihre hochkonservierte Struktur, erweisen sich ncRNAs als abbauresistent und fungieren auf diverse regulatorische Weisen. Trotz der schwach ausgeprägten Sequenzkonservierung, können homologe ncRNAs in anderen Spezies anhand von vergleichenden Methoden der Genomik identifiziert werden. Durch die allgegenwärtige Verfügbarkeit von Hochdurchsatz-Sequenzierungstechniken entstand ein reichhaltiger Datenschatz, der nun einen detaillierten Einblick in zelluläre Prozesse gewährt. Die Kombination aus genomischen und transkriptomischen Daten erlaubt ein besseres Verständnis von Mechanismen der Transkription, als auch eine funktionelle Charakterisierung einzelner Gene und deren evolutionären Herkunft. Eine analytische Prozessierung dieser riesigen Datenmengen ist dabei nur mit Hilfe optimierter bioinformatischer Algorithmen und Automatisierung einzelner Auswertungsschritte möglich.

Um die verschiedenen funktionellen Rollen von ncRNAs aufzuzeigen, beinhaltet diese Arbeit vier verwandte Studien, die sich mit dem Transkriptionsprozess aus unterschiedlichen Betrachtungswinkeln auseinandersetzen. Die erste Studie befasst sich mit einem ungewöhnlich langen untranslatiertem Bereich vor dem Division Cell Wall Gencluster in *Rhodobacter*, der eine regulierende ncRNA enthält. Im zweiten Teil wurde die enzymatische Prozessierung und der Abbau von 6S RNA in *Bacillus subtilis* anhand von Trankriptomdaten analysiert, um diesen finalen Teil des RNA Lebenszyklus zu rekonstruieren. 6S RNA ist dabei selbst ein globaler transkriptorischer Regulator und weit verbreitet in Bakterien. Anschließend wird der Fokus auf das eukaryotische System und die ubiquitäre RNase P gelenkt, die für tRNA Prozessierung zuständig ist. Anhand von Unterschieden im Aufbau durch eine optionale RNA und mehreren möglichen Proteinuntereinheiten in Organellen werden die phylogenetische Verbreitung und Charakteristiken innerhalb der eukaryotischen Evolution untersucht. Abschließend werden circRNAs betrachtet, einer Spezies von ncRNA die erst kürzlich Interesse durch ihr gehäuftes auftreten in Neuronen auf sich zog. Sie werden post-transkritional durch Back-Splicing erzeugt und konkurrieren daher mit der Wirtsgenexpression. Als Teil einer zoologischen Studie in sozialen Insekten wurden circRNAs zum ersten Mal in Bienen identifiziert. Das Ziel dabei war es tätigkeitsabhängige Unterschiede in der Expression von circRNA zwischen Ammen und Sammlern zu finden und damit mögliche Funktionen dieser ncRNAs zu bestimmen.

Erst die Kombination aus Methoden der Genomik mit transkriptomischen Daten ermöglicht in vielen Fällen eine funktionelle Analyse von ncRNA und erlaubt damit ein vielschichtiges Verständnis transkriptionaler Mechanismen. Es ist gelungen eine Riboswitch-ähnliche transkriptionelle Kontrolle durch UpsM nachzuweisen die einzigartig in Rhodobacteraceae ist. Außerdem zeigte sich, dass primär RNase J1 für die Maturierung von 6S RNA zuständig ist, während RNase Y den Abbau an internen Schleifen vorantreibt. Ein evolutionärer Vergleich der RNase P in eukaryotischen Organellen ergab, dass die RNA-Untereinheit teils stark unterschiedlich ausgeprägt und eine ausschließlich protein-basierte Variante potentiell durch horizontalen Gentransfer weitverbreitet ist. Mit circRNAs wurde eine ganze Gruppe an ncRNAs in dem sozialen

Modellorganismus der Biene charakterisiert und mindestens ein Gen zeigte signifikante Expressionsreduktion im Übergang vom Ammen- zum Sammlerstadium. Zusätzlich wurde ein überraschender Zusammenhang zwischen erhöhter DNA-Methylierung und RNA-Zirkularisierung gefunden. Die bioinformatischen Befunde in diesen Studien stellen eine Grundlage für weiterführende Experimente dar und zeigen gleichzeitig, dass wissenschaftliche Untersuchungen nie vollständig automatisiert werden können, sondern gründlicher Analyse mit teils angepassten Methoden bedürfen.

Acknowledgement

First and foremost I would like to thank my supervisor Dr. Marcus Lechner for the opportunity to conduct the research of this thesis. With his spontaneous pragmatism he was not only my experienced mentor during this time but became a dear friend to me.

Special thanks goes to Prof. Dr. Roland K. Hartmann who also made me part of his group and to whom I owe invaluable insights. He is also a supervisor of this thesis and enabled me to attend many scientific venues.

Next I would like to thank my PhD committee Prof. Dr. Dominik Heider and Prof. Dr. Alexander Goesmann.

I also thank the cooperation partners of the joint research projects, without whom the publication of my results would not have been possible. In particular Prof. Dr. Gabriele Klug and Dr. Markus Thamm who approached me with their projects that culminated in a scientific article each.

I am thankful for my time with the rest of research team. Particular mention is necessary for Sweetha Ganapathy and Dr. Markus Gößringer who proofread parts of my thesis. Best wishes also to the rest of my colleagues in no particular order:

Aileen, Dominik, Katja, Jana S., Nadine, Paul, Simone, Jana W., Laura, Arnold, Kerstin, Wiebke, Marietta, Clara, Isabell, Amri, Rebecca.

My parents supported my scientific career by always being there for me and encouraging me to ever new heights. Personally, I want to thank Aurora for supporting me along especially the last strides of my thesis and the inspiration in this phase of my life. Also I thank Lisa for what felt like a lifetime.

Contents

Glossary

BSJ	back-splicing junction
cDNA	complementary DNA
CDS	coding sequence
circRNA	circular RNA
ciRNA	circular intronic RNA
CpG	cytosine–phosphate–guanine
DCW	devision cell wall
DGE	differential gene expression
DNA	deoxyribonucleic acid
EIciRNA	exon-intron circular RNA
FDR	false discovery rate
JSR	junction-spanning read
MFE	minimum free energy
miRNA	micro RNA
mRNA	messenger RNA
ncRNA	non-coding RNA
NGS	next generation sequencing
nt	nucleotide
NTP	ribonucleoside triphosphate
ORF	open reading frame
P RNA	RNase P RNA
PRORP	protein-only RNase P
qPCR	quantitative polymerase chain reaction
RNA	ribonucleic acid
RNAP	RNA polymerase
RNase	ribonuclease
rRNA	ribosomal RNA
RT	reverse transcriptase
SCC	single cohort colony
sRNA	small RNA
TAP	tobacco acid pyrophosphatase
TEX	terminator 5'-phosphate-dependent exonuclease
tRNA	transfer RNA
TSS	transcription start site
UpsM	uspstream sRNA of *mraZ*
UTR	untranslated region

1 Introduction

The information flow from DNA to RNA molecules through transcription and on to proteins through translation makes up the foundation for the central dogma of molecular biology [1,2]. However, even besides additional DNA and RNA replication and reverse transcription from RNA to DNA [3] the picture is still much more complex than that. Only a portion of all genes encoded on chromosomal DNA contains a viable open reading frame (ORF) while a large population of transcripts consists of so-called non-coding RNAs (ncRNAs) [4]. An abundant amount of ncRNAs besides the prominent examples of ribosomal RNAs (rRNAs) and transfer RNAs (tRNAs) regulate the expression of other transcripts and are thus indispensable for fine-tuning many pathways (reviewed in [4]). These molecules are often defined by their well-conserved structure which is induced by internal base-pairing. As part of other transcripts (like riboswitches [5]), direct interference with other loci [6], or in interaction with the transcription machinery [7] ncRNAs can act as transcriptional regulators. Others possess catalytic activity, like the ribunuclease P (RNase P) which is in turn responsible for the processing of yet another ncRNA during tRNA maturation [8]. At the end of the RNA life cycle stands the degradation through nuclease activity.

The aim of this thesis is to investigate the impact of ncRNAs in transcriptional regulation and processing with bioinformatic methods. Following the layout of the central dogma,

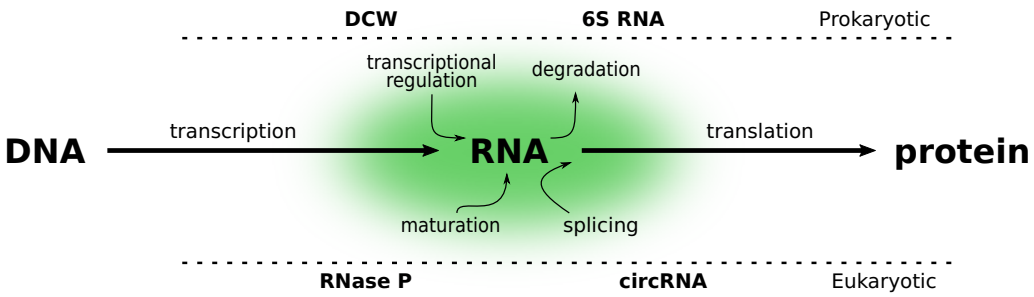


Figure 1: Schema of the thesis along the central dogma of molecular biology. Aspects around the RNA processing cycle are highlighted from four different perspectives: transcriptional regulation of the DCW in *Rhodobacter*, degradation of the two 6S RNAs in *Bacillus*, evolution of the tRNA processing ribozyme RNase P and splicing of circRNAs in honeybees.

the steps involved in the transcriptional pathway are illustrated on the basis of four transcriptomic and comparative genomics studies (two in bacteria and two in eukaryotes) in Figure 1. They exemplarily highlight the influence of a ncRNA in the transcription of the Division Cell Wall (DCW) cluster in *Rhodobacter* (Section 1.1.1), degradation of 6S RNA in *Bacillus subtilis* (Section 1.1.2), the evolution of the tRNA processing RNase P in eukaryotes (Section 1.1.3) and the biogenesis of circular RNA (circRNA) through post-transcriptional back-splicing (Section 1.1.4) and a link to genomic methylation (Section 1.1.5) in *Apis mellifera*.

In each of these studies, data from different sequencing approaches (detailed in Section 1.2 builds the foundation to understanding the characteristics of the investigated ncRNA and its role in the transcription/processing machinery. The key to processing the large amount of experimental data lies in the application of state-of-the-art bioinformatic algorithms (Section 1.2.4 & 1.3). However, with in-depth analysis beyond differential gene expression (DGE) custom tailored tools are necessary to interpret these results and follow up on the specific findings with advanced statistical genomics.

1.1 Biological Background

Among prokaryotes but especially among eukaryotes transcription of messenger RNAs (mRNAs) and ncRNAs is not as canonical as presented by the central dogma. A number of additional genomic features, such as epigenetics, enhancers and promoter sequences determine the initialization of transcription, while transcripts themselves are prone to degradation [9] or necessary maturation through ribonucleases. The bases of only one strand of the unwound DNA double-helix is complemented in 3'–5'-direction with an RNA transcript in 5'–3'-direction. With transcription termination, the RNA molecules dissociate from the transcription complex. In eukaryotes a methylated guanine is added to the 5'-triphosphate (forming a 5'-cap) of the transcript co-transcriptionally. The 3'-end is cleaved and adenines added as a poly(A) tail. Most transcripts are additionally segmented into multiple exons which are spliced together, while interspersed intronic sequences are spliced out to result in the mature transcript (detailed in Section 1.1.4).

Next, mature mRNA binds to the ribosome via particular binding sites and is translated into the encoded protein. Translation involves incorporation of one amino acid for each three nucleotides, or codon, that match the anticodon of a particular tRNA. Protein amino acid sequences are thus determined by their encoding mRNA or originating DNA segments. These usually are initiated by the typical start codon (`AUG` in eukaryotes, additionally `GUG` and `UUG` in some cases in prokaryotes [10]) and end with one of multiple stop codons making up an ORF.

However, RNA transcripts are usually longer than the coding sequence (CDS) of the ORF because they include a ribosomal binding site in the untranslated region (UTR) upstream of the start codon or contain terminator signals which halts transcription downstream of the stop codon. The term 'gene' refers to any functional hereditary unit encoded as part of the genome [11]. Even though the word is used very broadly in different contexts, genes usually include all sequence features necessary for the successful transcription and regulation of an RNA.

In the following, current research of various aspects of ncRNA processing and regulatory roles are being highlighted. These topics each frame one of the case studies that were investigated using a combination of next generation sequencing (NGS) and genomics in this work.

1.1.1 Division Cell Wall Cluster

Canonically, transcription of genes in bacteria depends on so-called promoter elements which consist of sequence motifs 10 and 35 nt upstream of the transcription start site (TSS) [12,13]. These promoters are distinctive for the governing sigma factors responsible for RNA polymerase (RNAP) holoenzyme binding [14]. While sigma factors like σ^{70} (*Escherichia coli*) or σ^{A} (its homolog in *Bacillus subtilis*) are associated with most of the housekeeping genes which are necessary for stable growing conditions, certain sigma factors primarily induce stress-related responses (reviewed in [15]). After transcription is initiated by polymerase holoenzyme binding, the unwound DNA strand is complemented until a termination signal is reached. This can either be a Rho factor binding site with a usually C-rich sequence after the end of a reading frame [16] or a Rho-independent hairpin structure (or intrinsic termination) in the nascent RNA caused by two consecutive stretches of high complementarity [17]. The detection of terminator structures will be outlined in Section 1.3.

Similarly riboswitches in non-coding regions upstream of an ORF can also block transcription due to high sequence complementarity [5]. However, this terminator-like signal can be turned off upon ligand binding which induces a structural reconfiguration and causes transcription to continue. This mechanism allows bacteria to react directly to metabolites without a highly complex signaling pathway.

Transcriptional Regulation of the DCW Clusters

Due to the compact size of bacterial genomes (compared to eukaryotes), genes are sometimes densely clustered in sequence on one strand with only one leading promoter region. These features are referred to as operons and their order is evolutionarily conserved to preserve expression efficiency. The genomic channeling hypothesis, states that genes in such a cluster are uniformly expressed and sequentially processed because their products work in conjunction to fulfill their biological function [18]. Besides operons for ribosomal genes and the *atp* cluster, the division cell wall (DCW) cluster (shown in Figure 2) is a prime example of a highly conserved gene cluster especially

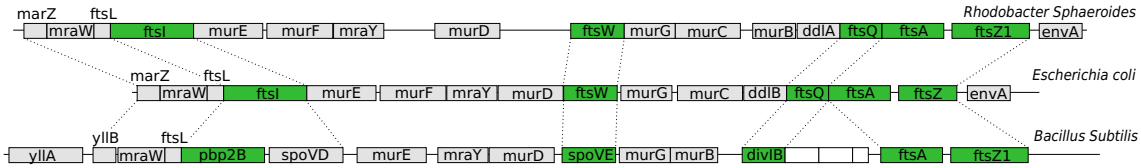


Figure 2: Conservation of gene sequence within the DCW cluster. The gene sequence in the DCW cluster of *R. sphaeroides* is compared to *E. coli* and *B. subtilis*. The black line represents the transcript of the entire operon, green boxes represent genes involved in septation, gray boxes represent genes involved cell wall synthesis and white boxes represent genes of unknown function. Adapted from [20]

in rod-shaped bacteria [19,20]. Individual genes encode either proteins involved in septation (green) or cell wall synthesis (gray) are tightly packed along the operon.

Along with the conservation of the genes themselves, their relative order within the cluster is also outstandingly conserved among diverse bacterial groups with the same cell morphology [21]. The transcriptional sequence of the entire cluster is thought to be essential for coordinating the cell wall division process during cytokinesis [18]. Due to the delicate timing after chromosomal segregation and spatial positioning of the dividing wall in the cell center, this basal process requires extreme efficiency and can be decisive for the cell's shape [22]. This is illustrated by the phylogenetic tree in Figure 3 which was computed by only regarding the relative order of genes within the cluster of the differently shaped prokaryotes [19]. In contrast to presence or absence of individual genes in the cluster, the DCW gene order correlates the cell shape as rod-shaped bacteria (green) along with hyphae-producing (blue) segregate distinctively from round (yellow), helical (red) and spirochaete (magenta).

While transcription of the cluster was thoroughly investigated in *E. coli* [20], it is yet unclear which impact multiple internal promoters and a 38 nt long 5'-UTR have on gene regulation and final protein ratios. In *Rhodobacter spheroides* the DCW cluster is lead by a 206 nt long ORF which was annotated as an orphan small RNA (sRNA) in a recent screening [23]. Transcription starting at the first promoter, *mraZ1p*, upstream of the *mraZ* gene can potentially continue up to the Rho independent terminator downstream of the last gene of the locus, *envA* [24]. Understanding the replication process partially guided by the DCW cluster in *Rhodobacter* is particularly interesting because this gram-negative bacterium is a model organism for photosynthesis [25]. Characterization of this sRNA and its regulatory function has not been assessed to this point and further investigation thus presents a case study in detailed experimental and bioinformatic analysis of transcriptional regulation.

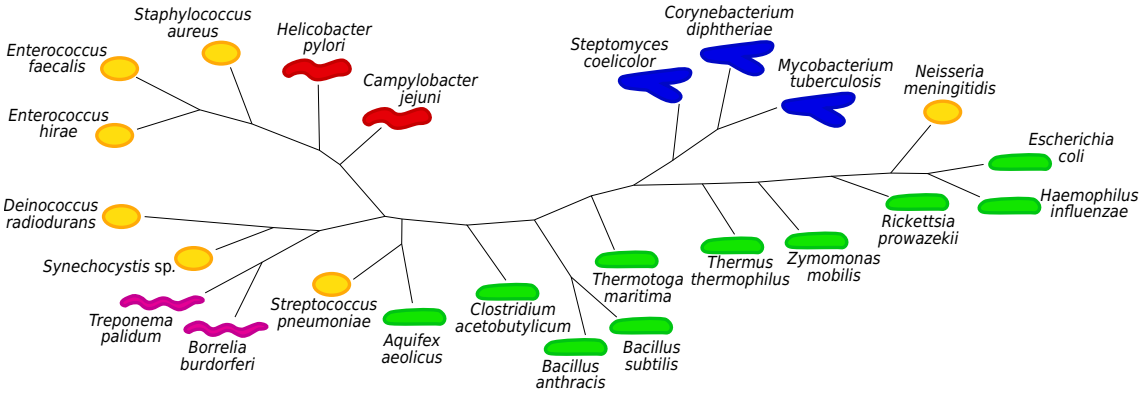


Figure 3: Comparison of bacterial shapes influenced by DCW gene order. The unrooted phylogenetic tree is based on the ordering of homologous genes within the DCW cluster. Rod-shaped cells in green, round in yellow, spirochaete in magenta, hyphae in blue and helix-shaped in red. Adaptation of [19].

1.1.2 6S RNA

Some ncRNAs possess well-conserved regulatory functions throughout the bacterial kingdom and 6S RNA is a prime example, as it was the first to be identified and sequenced in *E. coli* [26]. The RNA is approximately 200 nt long and highly structured with two main base-pairing stems along the centerfold and an internal bulge (visible in Figure 4). By resembling an open DNA promoter the central bulge [7] is able to specifically bind to the RNAP holoenzyme with the housekeeping sigma factor σ^{70} in *E. coli* or its homolog σ^{A} in *B. subtilis* and thus impacts global transcription regulation. However, knock-outs of the ubiquitous 6S RNA show only mild growth phenotypes during stationary phase of bacteria in culture. In *B. subtilis* a second paralogous 6S RNA (referred to as 6S-1 and 6S-2) was identified [27] that has an almost identical structure and can also be associated with RNAP. Conserved structures among Firmicutes containing two 6S RNAs are shown in Figure 4. Abundance in transcripts of both 6S RNAs varies between different growth phases [28]. While 6S-1 RNA reaches maximal intracellular expression levels in stationary phase induced by nutritional shortage, 6S-2 RNA is mostly expressed during exponential growth phases. This difference points to distinct physiological functions of the two paralogs. Both RNAs function as a template to short abortive product RNAs of 8–11 nt length [29]. Longer transcripts are produced with increasing NTP availability and induce refolding of the 6S RNA itself by base-pairing with the partially unwound terminal stem [30]. The conformational change in 6S RNA causes dissociation of the RNAP.

The majority of all transcripts (80 %) is prone to rapid decay with a half-life of fewer than seven minutes [9]. While bacterial mRNAs are swiftly translated often during transcription and by multiple ribosomes at once to avoid degradation by RNases, ncRNAs have to rely on their intrinsic structure that prevents them from attack. The 5'-end of 6S-2 RNA in *B. subtilis* is also known to be prone to degradation, as the first 11 nt are slowly cleaved off the ncRNA without further functional impairment [7]. Maturation of

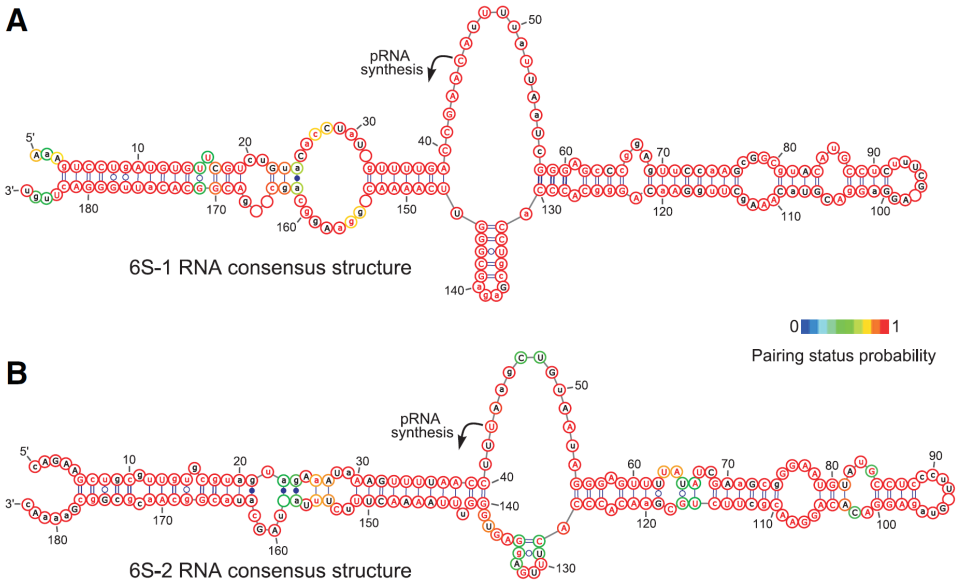


Figure 4: Consensus structure of 6S-1 & 6S-2 RNA in Firmicutes. Consensus structures are generated based on 14 6S-1 sequences and 16 6S-2 sequences from various species with two identified 6S RNAs. Original figure from [28].

the 6S RNA transcript seems therefore not be required. RNase activity is responsible for the degradation and maturation of RNA molecules but differs between different species [31,32]. There are 30 ribonucleases with different properties known throughout bacteria and the following are some of the major ones in *B. subtilis* relevant for this thesis:

- RNase P cleaves the 5'-leader of precursor tRNAs (detailed in the following Section 1.1.3)
- RNase PH and PNPase are 3'–5'-exoribonucleases which remove nucleotides phosphorolytically [33–35]
- RNase J1 is a 5'–3'-exoribonuclease which prefers monophosphorylated 5'-ends [36] but also exhibits endonucleolytic activity [38,39]
- RNase Y is an endoribonuclease which cleaves about 25 % of all mRNAs [40]
- RNase III is an endoribonuclease which cleaves double-stranded RNAs [41]
- RNase R and YhaM are 3'–5'-exoribonucleases which remove nucleotides hydrolytically [42,43]

1.1.3 RNase P

Another well-studied but also universally ubiquitous group of ncRNAs are tRNAs. These short and highly structured RNAs are required for translation of mRNAs into proteins in all forms of known life. Premature tRNA (pre-tRNA) transcripts have to be highly modified in order to fulfill their function [44]. Right after transcription, the 5'-leader of precursor tRNAs is removed by the endonuclease RNase P. Even though RNase P is also ubiquitous among cellular life, quite different architecture emerged independently throughout evolution to carry out this essential processing step in different organisms and organelles [45]. The ribozyme is made up of an essential RNA component (RNase P RNA or P RNA for short) and one additional protein as in the case of Bacteria or 4–10 in Archaea and nuclei of Eukarya [46,47]. These protein subunits promote substrate affinity to favor uncleaved pre-tRNAs over mature tRNAs which in turn increases catalytic efficiency of the holoenzyme [45]. While it could be shown that bacterial P RNA is capable of sustaining catalytic activity on its own under elevated salt conditions *in vitro* [8], the protein subunit increases catalytic activity especially under physiological salt conditions [48].

In all domains of life, P RNA is structurally highly conserved with up to 19 base-pairing regions (P1–19) [49] and contains five fairly conserved sequence motifs (CR-I–V) shown in Figure 5 [50]. CR-II and III are responsible for substrate specificity in bacteria, while CR-I and V are part of the catalytic center itself [50]. In eukaryotes a structurally similar ribonucleoprotein exists for mitochondrial RNA processing (hence RNase MRP), responsible for 5.8S rRNA maturation, which shares most associated protein subunits with RNase P and is evolutionarily related [47,51]. Both RNAs are highly dependent on their protein subunits to sustain catalytic efficiency even if P RNA is the essential catalytic moiety in RNase P [52,53].

Until recently, a plausible RNA subunit for mammalian mitochondrial RNase P was missing and researchers believed, that RNase P is imported to the mitochondrion to carry

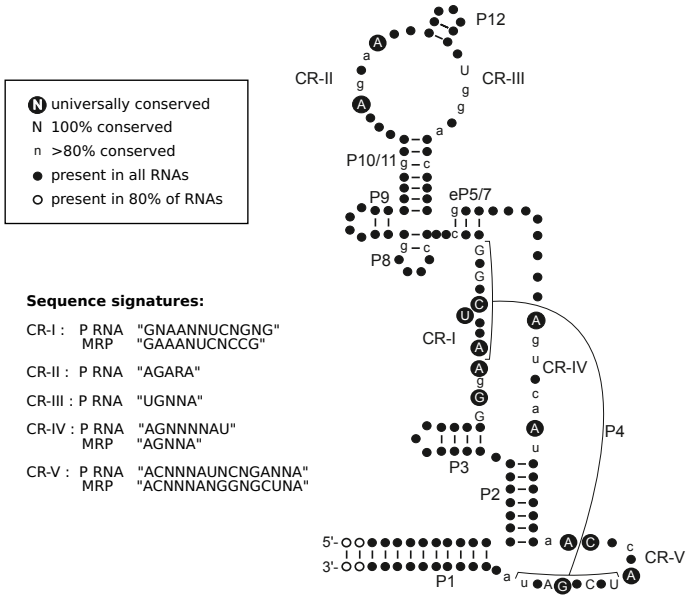


Figure 5: Consensus structure of RNase P RNA. Canonical structure of pairing regions P1–12 and most conserved sequence signatures CR-I–V. Adaptation of [49] with sequence signatures found in [50].

out tRNA maturation [54]. However, a protein-only RNase P (PRORP) exists in human mitochondria, which takes over the tRNA maturation step without the need of the central P RNA [55,56]. In mammalia mitochondrial RNase P consists of three mitochondrial RNase P proteins (MRPP1/MRPP2/MRPP3) that seem to have evolved separately from any of the known nuclear RNase P protein subunits [55].

Despite its strict functional conservation the various discovered forms of RNase P seem to have evolved into a set of quite distinctive machineries throughout eukaryotes. However, the exact evolution of organellar PRORP has not been investigated so far and P RNAs remain elusive in many already sequenced species. Thus comparative genomics approaches are required to understand this essential part of tRNA maturation in its evolutionary development.

1.1.4 Circular RNA

Eukaryotic genomics and transcriptomics differ further from prokaryotes in prolonged UTRs and often exhibit much longer gene sequences that also result in long transcripts before processing [57]. One of the reasons are elongated and more diverse promoters due to a multitude of transcription factors [58]. Additionally, most eukaryotic transcripts consist of multiple exons, separated by introns, which are removed to yield mature mRNA in a process called splicing. This results in a much more complex transcription and processing machinery which additionally has to be sensitive to splicing signals. The introns between exons thus contain the canonical splicing motif 5'-GT...AG-3' where splice factors are recruited [59]. Due to the added complexity of this process, alternative transcript variants can be derived through exon skipping or internal TSS through alternative splicing [60].

Only recently, another form of splicing was discovered between the 3'-end of exons and an upstream splice site [61]. This so-called back-splicing produces a looped single-stranded

transcript, termed circular RNA (circRNA). Even though the existence of circularized RNAs have been observed in the late 1970' as viroids [62] and in the cytoplasm of eukaryotic cells [63], this species of transcripts have long been overlooked in high throughput experiments. They were long believed to be an erroneous byproduct of the splicing process [64–66]. However, recent studies which showed that circRNAs are quite abundant and can be detected with high throughput sequencing sparked a new interest in the biogenesis and potential biological function of these transcripts.

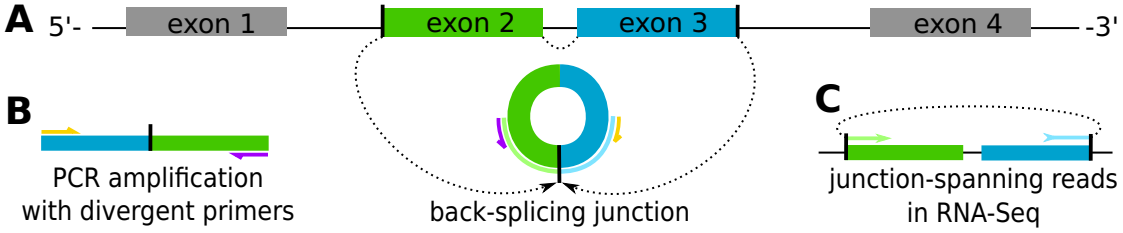


Figure 6: Circularization of exons through back-splicing. (A) Gene model along the transcript (black line) with exons (boxes) which are spliced during maturation. Exon circles resulting from back-splicing can be detected by either (B) PCR amplification with divergent primer pairs or (C) in RNA-Seq from junction spanning reads whose segments map chiastically to the reference.

Due to the lack of a translatable ORF, circRNAs can be considered a form of ncRNA. Because they do not possess a free 3'- or 5'-end, the circles prove to be resistant to degradation [67] which leads to their accumulation within the cytosol over time [63,68,69]. Circular transcripts are indeed much more stable (stable levels after 24 hours) than their linear counterparts, which possess a half-life of about 2 hours [70]. Additionally, circRNA do not always co-localize with their linear counterparts within the synaptic cytoplasm [71] and their expression can in some cases exceed that of their host gene (in some cases many fold) [72]. Back-spliced transcripts consist of a single or multiple exons from the coding region of the transcript but can also contain or be made up entirely by parts of the 5'- or 3'-UTRs [73]. While internal introns are usually excised from circRNAs [74], also purely intronic lariat forming circRNA exist, called ciRNA, which are generated during splicing, where the 2'-OH of an adenosine attacks the regular 5'-splice acceptor [75]. Lariat forming RNAs are quickly degraded by attacking the 2'–5' phosphodiester bond and were shown to perform poorly during reverse transcription, which makes them unlikely to be picked up by RNA-sequencing [76]. CircRNA containing exons and introns, called EIciRNA, are localized predominantly in the nucleus and have been shown to promote transcription of their parent gene by interacting with RNA polymerase II and U1 snRNP in some cases [77].

For the most common type of circular RNA, purely exonic circRNA, the flanking introns sequence, structure and length play a major role for general circularization efficiency [68,78–81]. Besides canonical splice signals present just up- and downstream of the circularized exons [79], flanking introns in human often contain Alu repeat elements that are complementary between 5'- and 3'-intron as in the case of *circ*ANRIL [72]. Mouse cells, generally lacking Alu elements (which are only conserved among primates), exhibit similar complementary regions that allow base-pairing and bring splice sites in close proximity of each other for back-splicing to occur [82]. Besides common short repeat elements, a 1.2 Kb long circularized exon of the *sry* gene is flanked by 50 Kb of

complementary introns in mouse testes [65,80]. Studies showed that introns flanking circRNAs in *Drosophila* [68] and *Danio rerio* [83] are not enriched with complementary repeat sequences that could allow for base-pairing, while other studies were able to find several prominent examples among the highest expressed circRNAs [73,79,81]. A circular splice product of the Muscleblind gene, *mbl*, in *Drosophila* is not surrounded by complementary sequences but rather littered with binding sites for its host gene's translated protein, suggesting that an interplay of protein binding and subsequent circularization creates an elaborate feedback loop for fine tuned gene expression regulation in this singular case [73]. Either way, there seems to be no congruent sequence motif throughout these complementary elements and even though many of the same exons are circularized across different species, base-pairing regions are not conserved even within species [80] making it hard to make out a universal mechanism.

In yet another case, a combination of complementary intronic sequences and binding of splicing factors were shown to regulate circularization of the *laccase2* gene also in *Drosophila* [81]. Without the complementary repeats in the flanking introns of *laccase2* circularization would not occur, however the splicing factors *hnRNP* and *SR* repressed circularization, effectively fine-tuning the ratio between linear and circular products. Targeted RNA interference of other parts of the splicing machinery in *Drosophila* cells showed that incomplete assembly or limiting of the spliceosome can lead to increased circularization of the pre-mRNA [70]. One hypothesis is thus, that circRNA are products of an inefficient or slowed pre-mRNA processing machinery by presenting a catalytically back-splicing alternative through, i.e., intronic base-pairing or RNA-binding proteins.

Potential Functions of circRNAs

Hitherto, a universal function of most circRNAs is yet nebulous, even though some potential functions particular to single circRNAs could be shown in the past. Above mentioned circRNAs of the Muscleblind gene in *Drosophila* are able to self-regulate its own transcription by binding to the protein encoded by its linear counterpart. Additionally, circRNAs from the *mbl* locus containing the endogenous start and stop codon were shown to be translated into peptides *in vivo* [84]. This however seems to be a unique example of a self-contained feedback loop.

A more widely debated function of circRNAs is their potential to act as so-called miRNA-sponges by carrying many conserved miRNA binding sites, being able to sequester corresponding micro RNAs (miRNAs) to counter-regulate their regulatory effect on the host gene. The prime example is the human CDR1as serving as a sponge for miR-7 with about 70 seed binding sites where the influence on miRNA abundance was measurably effected upon CDR1as over-expression [61,85]. Also circularized exons of *Sry* could be shown to compete for miR-138 binding [85] and *sickie* in *Drosophila* contains at least multiple miR-190 *in silico* predicted binding sites [68]. However, studies that screened circRNAs for global over-representation of miRNA binding sites come to the conclusion that conserved seed regions are not significantly enriched in circRNAs in general [71,76,86].

Honeybees as a Model Organism for Social Insects

A number of typical model organisms such as human and mouse [86] as well as *C. elegans* [78] , *D. rerio* [83] and *Drosophila* [68,73] have been investigated for the expression of circRNAs and are well characterized in this regard. However, since circRNAs are primarily traceable in neural tissue [68,71], we wanted to focus on the developmental and behavioral impacts of these transcripts in a relatively well studied social model organism. Previous to our research, honeybees (*Apis mellifera*) were suspected but not verified to harbor back-splicing mechanisms analogous to *Drosophila* and thus a promising research target [68,73,87]. Bees are organized in eusocial colonies where they exhibit caste-like system divided male drones, a queen and female workers [88]. The workers can be classified into further task-related groups of nurses, that remain inside the hive and care for the brood and foragers that leave the hive in order to collect pollen and nectar. A worker bee will typically start to perform nursing activity at day 3–12 and assume foraging tasks after approximately 20 days [89]. With this shift in roles also comes a dramatic change in environment, as the bees grew up in the dark, relatively small hive and now have to be able to rely much more on visual than tactile input in order to orientate themselves outside [90]. Through human intervention (hive trickery) of removing a large part of the hive population, a part of the returning foragers shows a remarkable plasticity in behavior by reverting into the nurse role in order for the colony to survive [91,92]. The worker bees in such a single cohort colony (SCC) then have all the same age and can be used to rule out age-related effects in experiments. Honeybees thus present an optimal model organism to study a task-related differential expression of circRNAs without a confounding age bias.

1.1.5 DNA Methylation

Epigenetic alterations of the DNA have a profound influence on subsequent gene transcription. Methylation, thus the addition of a methyl group at the C5 position of the benzene in a cytosine (5mC) represents the most common and best-studied modification at the genome level. Especially cytosines followed by a guanosine (CpG) are often found to be methylated throughout eukaryotic exonic sequences, allowing for a symmetrical methylation on the opposite strand of the `G:C` base pair [93]. However, spontaneous deamination of 5-methylcytosine can lead to a conversion to thymine resulting in a `T:G` mismatch that is either repaired to the original `C:G` or erroneously into a `T:A` pair (effectively introducing a `C->T` mutation) [94]. Especially in eukaryotes this conversion, over time, lead to regions of drastic under-representation of cytosines where no evolutionary pressure acted against this kind of mutation [95]. Stretches of mostly CpG repeats – dubbed CpG islands – in mammals can be found primarily in gene promoter sequences or within the gene body as part of alternative promoters [96]. These regions are thus under elevated evolutionary pressure counteracting mutagenesis. Additionally, methylation in these regions could be actively removed by DNA demethylase [97] or converted by Tet family enzymes to 5-hydroxymethylcytosine (5hmC) which is, in turn, a demethylation intermediate [98].

Methylation especially of CpGs in the promoter region results in efficient silencing of the downstream gene's expression [99]. Thus the global methylation status changes during development of, i.e., bees and ants from fertilized egg, larvae, worker and queen [92]. Some invertebrates such as *Drosophila* and *C. elegans* exhibit little to no methylation similar to *Saccharomyces* due to a lack of DNA methyl-transferase 3 (Dnmt3) responsible primarily for most of the *de novo* methylation especially during development. *Drosophila* possesses only Dnmt2 which is responsible for $tRNA^{ASP}$ methylation [100]. As a result, CpG frequencies are as statistically expected throughout the genome in these species. The western honeybee *Apis mellifera*, on the other hand, shows differentially methylated regions within certain genes comparing different castes to totipotent female eggs [101]. Researchers were able to initiate queen development of larva through nutritionally induced RNA interference of Dnmt3 which leads to a globally decreased methylation status [102]. Despite this finding, queens and worker bees do not seem to exhibit differentially methylated regions immediately after emergence from pupal stage [92]. The same study showed evidence of reversible methylation patterns during the transition from nurse to forager tasks in adult bee brains. Methylation in *Apis mellifera* takes place mainly within the gene body as opposed to the promoter region in vertebrates and in some cases induces alternative splicing patterns [92,103,104]. It is therefore possible, that changes in the methylation status of certain exons induce alternative transcription initiation tightly coupled with circularization of relatively short aberrant transcripts.

1.2 Sequencing

Even though experiments with individual genes can be conducted much more thoroughly and accurately based on quantitative polymerase chain reaction (qPCR) and northern blots, high throughput sequencing has become a powerful approach to many molecular biology problems. Due to improvements in throughput along with competitive pricing, global transcriptomic analyses are a *de facto* gold standard to investigate gene expression and regulatory impacts. The following section is dedicated to the rise of different sequencing techniques and their applications to particular molecular biology experiments.

1.2.1 Techniques

One of the first feasible methods of sequencing DNA fragments was presented by Fred Sanger [105]. This first generation technique uses a short DNA primer fragment that is annealed to complementary DNA regions of interest which had previously been denatured by heat. Starting at this new double-stranded position DNA polymerase is able to complement the entire strand in the presence of an excess of all four deoxyribonucleoside triphosphates (dNTPs: dATP, dCTP, dGTP, dTTP). The process is done in four parallel solutions, each containing one additionally modified di-deoxyribonucleoside triphosphate (ddNTP) which upon incorporation into the sequence leads to termination of further elongation of the particular molecule. Due to the direct competition of dNTPs and ddNTPs in different ratios during elongation, all strands are stochastically complemented to different positions. Resulting different length products can be distinguished by electrophoresis on a polyacrylamide gel for each base individually and the sequence can

be deduced from the combination of the four bases [106]. The approach is quite time consuming, even though the size limitation of gels could be circumvented by combining multiple gels [107,108]. The major shortfall of this technique is that a larger DNA molecule had to be deciphered sequentially in a process called 'primer-walking', where each starting primer is taken from the end of the last sequencing iteration [109].

The first effort to parallelize the process for longer DNA sequences was introduced with the idea of shotgun sequencing, where the DNA molecules are fragmented virtually randomly (thus like by shotgun blast) and random initialization primers could be used [110]. As a result, multiple researchers could sequence the DNA in parallel and would have to combine their sequences afterward in order to reconstruct the total sequence by overlapping regions. Quickly the task of assembling the large amount of short sequences fragments (or reads) into contiguous stretches (or contigs) by hand became infeasible for large projects. Instead, computers which became widely available during this time in academia were programmed to perform the assembly [111].

Next Generation Sequencing

The next major disruption of molecular biology came about with the invention of NGS techniques based on the 'sequencing-by-synthesis' approach in picolitre reactors pioneered by 454 Life Science [112]. The prominent idea of these next generation methods is the massive parallelization base calling through pyrosequencing of complex libraries of short random fragments (outlined in [113]). In Illumina sequencing (outlined in Figure 7) specific sequencing adapters are ligated to the double-stranded DNA fragments (A) which scatter and attach to the solid phase of the flow cell surface randomly (B). Similarly to previous approaches DNA polymerases are used, but now in two steps. First to multiply single sparsely distributed fragments through bridge-PCR amplification to neighboring free adapters (C) to form uniform PCR clusters (D) and secondly by elongating all clonal fragments in one cluster with 3'-blocked NTPs simultaneously during the sequencing step (E). Elongation is halted after each nucleotide incorporation by the 3'-terminal block. The newly incorporated NTP is coupled to a fluorescent dye which can be detected under ultra-violet light [114] by a camera (F) and the dye is subsequently removed together with the 3'-block in order to initiate the next cycle. Due to the proximal distance of fluorescent

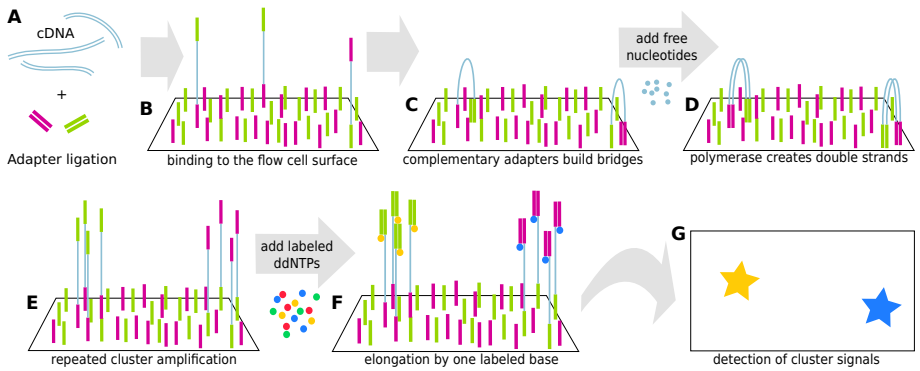


Figure 7: Illumina sequencing scheme. (A) Adapters are ligated to the cDNA fragments and (B) bound to the flow cell. (C) Adapters build bridges nearby and (D) a second strand is synthesized. (E) Amplified PCR clusters (F) are elongated with labeled ddNTPs. (G) Fluorescent clusters are recorded to induce the sequence of the molecule.

molecules within the densely populated PCR clusters, which are fixated to the surface of a flow cell, a clear spot for each cluster in different colors is produced each cycle. The readout of each cluster is then interpreted over time as the sequence of the underlying DNA fragments [115].

Transcriptomics

Because RNA transcripts can be re-transcribed into complementary DNA (cDNA) sequences through reverse transcriptases (RT) found in and isolated from retroviruses [3], researchers are equally able to sequence expressed gene transcripts indirectly [116]. This type of sequencing is commonly referred to as RNA-Seq but entails host of possible protocols for specific experimental applications [117]. While *de novo* sequenced genomes first have to be assembled into continuous sequences and entire chromosomes later on [118], mapping transcriptomic reads onto a previously sequenced and fully annotated reference genome is much less complex and can thus be done more quickly for more in-depth analysis [113]. Thankfully, the recent count of most studied organisms is at 81,345 reference genomes as per release 89 (July 2018) of the NCBI RefSeq project. Thus, quick transcriptional analysis of most laboratory strains is possible in a reasonable time and with affordable hardware [117].

1.2.2 Library Preparation

While whole-genome sequencing aims to deliver a complete chromosomic overview by producing reads covering the entire genome evenly [118], transcriptomics focuses on quantitative differences on a transcript level [117]. In order to prevent DNA templated reads during PCR, cell extracts are thus treated with DNase after lysis, leaving only RNA molecules after purification [119]. Even though rRNA is by far the most abundant class of transcripts in cells, for most RNA-Seq experiments it is not of interest [120]. It can be diminished by targeted rRNA depletion for the particular species with complementary oligomers available in commercial ribo(-) kits. In eukaryotes mRNAs possess a poly(A) tail for further processing signaling and can easily be selectively enriched by capturing them with complementary poly(T) oligomers [117]. Additionally, if particularly ncRNAs are of interest, size exclusion chromatography can be applied to a sample to select tRNAs (~90 nt) [121], bacterial sRNAs (50–500 nt) [122] or eukaryotic small nuclear RNAs (snRNA, ~100 nt) and miRNAs (~22 nt) [123], which are smaller than most mRNAs. Longer transcripts exceed the length limitation of most modern high throughput sequencing methods with read lengths between 30–400 nt and have to be fragmented into smaller RNA molecules using hydrolysis or after reverse transcription by sonication, mechanical shearing or restriction enzyme fragmentation of cDNA molecules [124]. Both methods as well as other library processing steps induce further biases in gene coverage of base-paired regions or towards one of the transcript ends [119]. In order to capture the original length of especially mRNA transcripts, paired-end sequencing can be employed in library preparation to sequence transcripts from both ends [118,125]. Doubling the effective read length per transcript might not be enough to cover the entire sequence but the original length can be deduced after mapping to a reference.

Enrichment of Primary Transcripts

Transcripts are often processed before translation and can be prone to degradation by RNases, leaving partially degraded transcripts in RNA libraries. Primary transcripts are lead by a 5'-triphosphate end in bacteria that is added at transcription initiation. Terminator 5'-phosphate-dependent exonuclease (TEX) degrades RNAs with a 5'-monophosphate end which are a product of ribonuclease degradation. It thus effectively enriches primary transcripts among the RNA sample [126]. For reverse transcription into cDNA for library preparation 5'-triphosphates are converted to monophosphates by tobacco acid pyrophosphatase (TAP) [127]. Relative increases in read coverage after TEX/TAP treatment are therefore used for TSS detection [128].

Transcript degradation through different exonucleases can also be explored in specific knock-out strains. Various exoribonucleases degrade transcript either from an unstructured 3'- or 5'-end, while endoribonucleases cleave either single- or double-stranded regions of RNAs. A differential analysis of the transcriptome in RNase knock-out strains compared to wild-type conditions can detect transcriptional variants (longer UTR) and structural features.

Enrichment of Circular Transcripts

RNase R is a 3'–5'-exonuclease that unspecifically degrades most unstructured RNA molecules. It is used to enrich circular RNAs (circRNAs) because they do not possess a free 3'-end. Preparation methods of modern high-throughput sequencing protocols often select against circRNA because they do not possess a 5'-triphosphate cap or poly(A)-tail and are thus not picked up during usually performed amplification steps [126,129]. However, targeted enrichment of cirRNA through RNase R treatment, a magnesium-dependent 3'–5'-exoribonuclease, will digest most unstructured linear transcripts and is thus suitable for preparation of sequencing libraries and qPCRs to study circRNAs [67]. Optimally, linear transcripts carry at least a 5 nt unpaired overhang at the 3'-end for the nuclease to attack.

DNA Methylated Detection with Sodium Bisulfite

Besides detecting methylations indirectly by capturing them with antibodies or immunoprecipitation with methyl-CpG binding domain harboring proteins [130,131] the gold standard is nowadays treatment with sodium bisulfite to convert unmethylated cytosines into uracils [132]. The low abundance of methylated cytosines will be unchanged by the treatment and can be detected during sequencing as the only remaining cytosines while the converted uracils are replicated as thymines during PCR amplification. A similar modification of cytosines, 5hmC, is abundant in some cell types and will react with bisulfite to form cytosine 5-methylenesulfonate which is also detected as a normal cytosine during sequencing [133]. Therefore the two modifications are indistinguishable to this method. Bisulfite genome sequencing provides information about the methylation patterns of single molecules and thus allow an in-depth analysis of actual DNA methylation in a cell.

1.2.3 Quality Processing

The imaging data captured from the fluorescently labeled nucleotides during synthesis is standardly converted into a sequence of base calls for each cluster. The produced FASTQ formatted output does not only contain information about the most likely base sequence and originating cluster coordinates for documentation purposes, but also the associated quality of each base [134]. Nowadays the universally adopted PHRED score [115,135] is used to compute a numerical value in relation to the estimated probability of a base calling error, P_e, and given for each nucleotide encoded as a corresponding ASCII character in the FASTQ file [134]:

$$Q_{\text{PHRED}} = -10 \quad log_{10}(P_e) \tag{1}$$

Thus, it is possible to interpret ambiguous color signals due to heterogeneous NTP incorporation after sequencing. Different sequencing platforms and preparation methods can result in specific quality reductions and sequence biases particularly in the start or towards the end of a read [119]. In order to assess the quality of a sequencing run based on the resulting FASTQ data FastQC [136] is an invaluable tool that quickly produces visual representations of read lengths, repeating sequences, biases and potential contaminations. Regardless of the inspection, reads should be "cleaned" before the mapping step to trim reads with low-quality bases or ligated adapters at either end to maximize the mappability of the remaining high-quality fragment. Read-trimmers like trimmomatic [137] use a sliding window approach to remove trailing sequences with subthreshold quality. Reduction of base quality towards the end of many reads can be a result of unspecific synthesis beyond the actual transcript length or heterogeneous PCR clusters. When working with paired-end reads, processing of both mates simultaneously can ensure complementarity in low-quality regions or lead to the rejection of both in cases of contradictions. Additionally, the tool filters known contaminations, such as sequencing adapters and PCR primers.

1.2.4 Mapping Algorithms

Before modern read mapping algorithms for NGS were available, BLAST [138] was the most widely used software for the attribution of sequencing reads [139]. At its inception, it allowed incomparably fast alignments of unique sequence fragments of medium to long length to a database of many hundred genomes at once. With improvements in sequencing methods the number of reads quickly exceeded a lookup of entire libraries in a feasible time frame. Moreover, NGS reads often contain too many errors to be aligned unambiguously by BLAST. These errors arise either from sequencing artifacts that survive quality trimming or represent genuine deviances from the reference sequence. Despite many improvements implemented on top of BLAST [140–144] that exist today, dedicated mapping programs are able to process large libraries much more rapidly [139]. This new breed of specialized short read mapping algorithms accounts for more mismatches and gaps in the alignments to increase sensitivity while further improving processing speeds by applying heuristics and improved lookup techniques.

While initial approaches, such as SOAP [145] and MAQ [146] relied on string hashing, a next generation of algorithms, such as Bowtie [147], BWA [148] and SOAP2 [149], reduced memory usage by leveraging the Burrows-Wheeler-Transformation [150] for reduced sequence representation. The transformation uses a reversible cyclic permutation of arbitrary character strings which groups identical symbols. Such rearranged sequences can easily be compressed and therefore processed faster and with reduced memory requirements. In order to make implementations of this transformation for RNA-Seq data feasible in linear time, these approaches use suffix trees to compute all possible suffixes for a given genome database. Short read sequences are then traversed through this tree with possible jumps, or suffix links, for mismatches or gaps in order to determine a possible originating coordinate of the read within the reference. Suffix trees can be represented as suffix arrays, a compact data structure without the so-called suffix links which account for matching errors. A recent algorithmic approach using enhanced suffix arrays with additional tables to account for alignment errors is implemented in the widely used short read mapper segemehl [151].

All up-to-date short read mappers will output in standardized SAM plain text file format, which contains one line per matched read segment with information about coordinate, strand, mismatches and possible read pair information relative to the reference [152]. This independently developed file format assures inter-operability with downstream analysis tools and enables to flexibly create specialized analysis pipelines for any experimental setup. Even though plain text files are convenient to work with, the suboptimal encoding is extremely storage intensive. Hence, most tools support lossless, compressed BAM binary files or the data can be encoded and decoded on the fly to save hard disk space [152]. Plain text and compressed files can easily be filtered and manipulated (in stream) on the command line via samtools.

Split-read Mapping for Splice-junction Discovery

In newer iterations, segemehl implements a feature which allows for split-read mapping, where partial segments of the read can be mapped individually to account for cis- and trans-splicing [153]. Bowtie2 [154] also allows for mapping of spliced reads with the additional TopHat2 [155] software which are both part of the Tuxedo pipepline [156]. Novel algorithms with similar implementations like MapSplice [157], SpliceMap [158] and STAR [159] emerged, dedicated to annotation and quantification of *de novo* detected splice sites. The latter also implements enhanced suffix arrays similar to segemehl in order to rapidly find, cluster and stitch together short seed matches. All of these programs have their advantages as well as disadvantages and should therefore be chosen depending on individual requirements (benchmarked in [160]). Generally, segemehl and STAR map the most reads with particularly high sensitivity. Both require enormous amounts of memory on the processing computer (up to 70 GB for the human genome [151]) which is often still infeasible for everyday usage in most labs. While segemehl has relatively high accuracy for transcriptomic data, BWA and STAR both are known to produce fairly high false positive rates but map reads in a fraction of the time. Bowtie2 ranges in the middle of most of those properties and thus also widely used.

With the recent surge in interest in circRNAs, a multitude of novel software for the detection of circRNAs in RNA-Seq data has emerged and many established mapping algorithms have been adapted to cope with split-read mapping also in a chiastic order to account for back-splicing. Reads which span the back-splicing junction (BSJ) of a circRNA will be split and the first half mapped downstream of the second read segment in the transcript. RNA-Seq is not regarded as ideal for accurate validation of genuine circRNAs alone because chimeric reads can also spring from exon scrambling during splicing [64] or genomic exon shuffling [161]. However, studies showed that circRNAs with strong support of many junction spanning reads (JSRs) in RNA-Seq could be reliably verified with other methods [68,73]. The abundance of this species of transcripts is marginal next to highly expressed linear transcripts. Thus, new library preparation protocols enable improved yield of JSRs through rRNA depletion and RNase R treatment [68,76,162].

Current algorithmic approaches to the identification of BSJs all have different advantages and shortfalls while they will all detect the same high confidence and strongly expressed circRNAs from the same data [163,164]. Major differences lie in the particular mapping algorithm used because these largely determine sensitivity, computation time and memory requirements. Initial approaches like `find_circ` [61] and `circExplorer` [165] relied on mapping the data twice to sieve out linear transcripts in the first step and create anchor points for JSRs. In a second step remaining reads are then mapped against potential junctions. In contrast to these time consuming two-phase approaches, `segemehl` [153] manages to detect JSRs in one mapping step and is thus much simpler to use. The accompanying tool `testrealign` filters split reads that occur at conventional splice sites are due to trans-splicing events and it does also report split-reads that are mapped chiastically as produced by BSJs. Additionally, taking annotations and splice signals in the sequence into account, `CIRI` [166] presents another source of high certainty circRNA detection leveraging the very fast `BWA` mapping algorithm [148]. While initial versions relied exclusively on well annotated references, the newer version, `CIRI2` [167] is also able to handle *de novo* identification and together with `segemehl` ranks well in latest circRNA detection benchmarks [163,164].

Methylation Detection

Another common application of mapping algorithms is the detection and quantification of methylation sites on a genome level that can be explored experimentally by treating DNA with sodium bisulfite [132]. The treatment converts unmethylated cytosines into uridines which are replaced by tyrosines during the PCR step of the sequencing procedure. While methylated cytosines are not converted, they represent the scarce minority and thus sequenced reads are practically void of `C`s. Mapping reads with so many deviances from the reference genome becomes infeasible especially for `GC`-rich regions. Therefore, most mappers implement a mode that converts the reference genome into a `A-G-T` only alphabet, before the subsequent mapping step with decreased specificity parameters [154,168,169]. Single, rarely occurring methylated cytosines are mapped as mismatching `C`s to the reference and are then evaluated by downstream analysis tools like `bismark` [169]. The software then outputs absolute and relative methylation status for each sequenced cytosine in the reference to infer either region-specific accumulation of

methylation or backtrack individual methylation patterns on a per-read basis. However, this method cannot distinguish between 5mC and 5hmC [133].

1.2.5 Differential Gene Expression

Comparing transcriptomic libraries of sufficient sequencing depth allows for a comprehensive statistical analysis of quantitative differences in gene expression between experimental setups [128]. Quickly a whole field of research formed around the topic of DGE which takes the number of reads uniquely mapped to a transcript into account. Unambiguously attributing mapped reads to the single gene is non-trivial in the case of overlapping genomic loci or reads mapping to the opposite strand due to strand-unspecific sequencing protocols [170]. These issues are addressed by specialized algorithms like `featureCounts` [171] by rapidly counting only non-ambiguous reads of many libraries at once. An even more advanced software like `kallisto` [172] is available which skips the entire mapping process and instead counts the number of reads with a sufficient number of short sub-sequences (k-mers). However, this approach does not allow for an in-depth analysis of read mapping positions required for the experiments presented in this thesis.
Before a statistical DGE analysis can be carried out, the read counts of the different libraries have to be normalized in order to compensate for differences in RNA yield, flow cell occupation or sequencing depth [119]. Now, specialized packages like `DESeq2` [173] can be used to test for significantly up- or down-regulated genes by creating a negative binomial distribution of expression variance depending on the mean expression of each gene. Because every gene is tested individually for significant variance between conditions, probabilities are corrected for multiple testing with the Benjamini-Hochberg false discovery rate (FDR) method [174].

1.3 RNA Bioinformatics *in silico*

Besides with transcriptomics, ncRNAs can well be studied and characterized with other predictive bioinformatic methods *in silico*. Left untranslated these transcripts are mostly defined by their tertiary structure determined by general folding constraints and possess important regulatory and post-transcriptional functions [175]. Additionally, homologous ncRNAs can be compared between different organisms similar to mRNA and protein sequences, even though sequence conservation is rather poor in these cases [176].

1.3.1 Sequence

Conventionally, mRNAs or proteins can be easily compared among related species because their sequence is often well-conserved. Alignments of two or more similar sequences can be arranged as to maximize the number of matching bases or amino acids. For each pairwise alignment a similarity score is calculated derived from producing the maximal number of k-tupels [177]. One of the most prominent examples of advanced multiple sequencing alignment algorithms are `ClustalW` [178] or its successor `ClustalOmega` [179]. They will introduce gaps (`-`) into a sequence where a deletion took place (or an insertion from the point of view of the other sequence).

1.3.2 Structure

Because ncRNAs do not encode for a certain protein sequence, their sequence in turn is hardly conserved. In order to annotate a homolog of a known ncRNA in somewhat related species a simultaneous comparison between sequence and structure has to be employed. The `infernal` package uses hidden Markov model for comparison of the sequence to an alignment of known species and covariance models of a structural alignment of the same references in parallel [180]. Structural alignments can be computed with `RNAalifold` or `LocaRNA` [181], which also take a consensus structure of multiple sequences into account. Alternatively, a collection of well-curated covariance models for known ncRNAs is available in the so-called `Rfam` database [182] similar to protein family database `Pfam`. Annotation of novel structural RNAs should be guided by phylogenetic sampling of transcriptomic data from familiar organism to discern functional ncRNAs from transcriptional noise [183].

`RNAfold` and similar tools from the `Vienna RNA` package have become the *de facto* default for *in silico* RNA structure and folding energy prediction [184]. It efficiently explores the folding landscape of possible base-pairs in order to minimize the free energy (MFE) of the molecule based on empirically deduced energy constrains. Downstream tools, like the combination of `RNApfold` plus `barriers` [184] and `treekin` [185], are able to leverage this approach to explore different substructures possible during the transcriptional process. It is thus possible to discover alternative structures a given RNA molecule can produce depending on external parameters, e.g., as in riboswitches. A slightly different approach and energy parameters are used by `mfold` and the tools can thus be used to provide complementary predictions especially on regional substructures [186].

Other tools are specialized to predict the interaction of two RNA molecules to predict hybridization with `RNAcofold` [184] and `RNAhybrid` [187] or reflect the accessibility of a target mRNA for a siRNA by accounting for the unfolding of the target sequence with `RNAup` [184], `CopraRNA` [188] and `IntaRNA` [189]. Some of these tools can also be used to screen for possible interactions genome wide. Similarly, `transtermHP` [17] predicts bacterial terminator hairpins based on strong local complementarity which induce strong dissociation of the RNA polymerase downstream of genes or regulates gene expression in upstream regions [190]. Additionally to the often completely complementary hairpin stem, bacterial terminators are frequently lead by an `A`-rich 5'- and trailed by a `T`-rich 3'-flank also scored by the algorithm.

Even though most of these tools are older than modern sequencing techniques, they are irreplaceable for in-depth analysis of transcriptomic findings beyond the canonical gene expression. Using these specialized *in silico* methods for sequence and structural prediction purposes sheds light on the mechanistic processes of transcription itself.

1.3.3 Phylogeny

Because different species developed from common ancestors through evolution over time, some genes remain comparable among related species if they still fulfill a similar function. These genes are considered orthologs. Over time, however, random mutations within the sequence may change the gene and its function, especially when the

evolutionary pressure is relieved due to a duplication event or changing environmental conditions which render the gene non-essential. Two genes that are the result of such a duplication event are considered paralogous to each other. Both are examples of homologous genes that derive from a common ancestor and can be compared in sequence alignments or based on their structure. Reversely considering the similarity of genes in different species, one can deduce their likely evolutionary distance based on assumed mutation rates. Phylogeny then is the discipline of evolutionary relationships between species based on single or multiple hereditary traits.

A phylogenetic assessment can be made based on sequence alignments of, e.g. , protein sequences. Based on the alignments the distance between the sequences is computed and a phylogenetic tree can be generated, e.g. in the simplest way via neighbor joining. This method is implemented in `ClustalX` [178] and iteratively joins the closest matching sequences based on their distance into a branch until the entire tree is populated. Computing the maximum likelihood estimate for a given alignment is another approach to generating a phylogenetic tree considering each position as individual parameters. The software `RAxML` [191] implements this concept by finding the tree that best explains the observed sequence differences under the assumption of a specific substitution and rate change model. Various such models which are based on substitution frequencies observed in nature can be used and are represented in amino acid substitution matrices such as `BLOSOM` [192] and `LG` [193]. Choosing the best fitting model can be automated by comparing them with observed frequencies in the alignment itself using tools like `prottest` [194]. In order to assess the stability of each branch in such a final tree, a subset of the tree nodes can be used to re-create the different branches in a statistical sampling method called "bootstrapping" [195] which is also implemented in `RAxML` [191]. A bootstrapping value represents the portion of subtrees in the samples agreeing with the original. One is assigned to each branch point to reflect how many of the re-sampled trees contained the identical relationship between the two emanating branches.

2 Methods

This chapter elaborates on the methodological contributions to four different projects covered by this thesis. As long as not indicated otherwise, analyses were carried out entirely by myself or in cooperation with at least equivalent participation. Due to the *in silico* focus of my work mostly algorithms with version number and parameters (when deviating from defaults) will be listed for each individual project. Software was installed and executed locally on different computers running `Linux Mint 17 Qiana (GNU/Linux 3.19.0-80-generic x86_64)`. In some cases publicly available datasets from other research groups were included in the analysis. These are stated with their respective database ID to make backtracking of the original data and study parameters possible. Genome sequences, if not indicated otherwise, were obtained from NCBI RefSeq database for the respective representative genome.

2.1 DCW

We aimed to characterize a ncRNA in the DCW cluster UTR by comparative transcriptomics and investigated its potential structurally induced regulatory role for subsequent gene expression based on RNA folding predictions.

2.1.1 Transcriptomics of the DCW UTR

To assess phylogenetic conservation of an extended 5'-UTR, the upstream region of the *mraZ* gene was compared to that of other species were it leads the DCW cluster. Sequencing data was obtained from the NCBI SRA database [196] (accessed 8/10/2015) for closely related bacteria and combined in-house data in order to annotate TSSs within 300 nt upstream of the *mraZ* locus. Especially valuable are TEX treated libraries where available in order to discern primary transcripts from processing artifacts. An overview of used libraries and reference genomes can be found in Table 1. Reads from these libraries were mapped to the respective reference genomes using `segemehl v0.2.0-418` [151] with default parameters after quality trimming using `Trimmomatic v0.33` [137]

with a quality threshold PHRED score of 25 in a sliding window of size 3 bases. Only reads of size > 14 nt after trimming were considered to minimize the influence of remaining partially processed transcripts as primary transcripts should be considerably longer. Rho-independent terminators in the 5'-region up to 300 nt upstream of the annotated *mraZ* gene were predicted using `TranTermHP v2.09` [17]. Only the highest scoring hairpin structure with an MFE below -10 kcal/mol according to `RNAfold` [184] was considered a genuine terminator.

Table 1: Transcriptomic data for TSS determination. Overview of RNA-Seq libraries and treatment.

Species	RefSeq ID	Data source	Treatment
Rhodobacter sphaeroides	NC_007493.2	SRR2149464	TEX
Rhodobacter capsulatus	NC_014034.1	SRR4244374	TEX
Ruegeria pomeroyi	NC_003911.12	SRR1823766	.
Phaeobacter gallaeciensis	NC_023137.1	SRR1798598	.
Dinoroseobacter shibae	NC_009952.1	SRR1177020	.
Caulobacter crescentus	NC_011916.1	SRR1273068	.
Sinorhizobium meliloti	NC_020528.1	SRR701993	size exclusion (>200nt)
Myxococcus xanthus	NC_008095.1	SRR882104	.
Bacillus subtilis	NC_000964.3	unpublished	.
Pseudomonas aeruginosa	NC_009656.1	SRR2174566	.
Salmonella enerica	NC_011083.1	SRR1532984	.
Escherichia coli	NC_000913.3	[197]	.

Transcribed sRNAs leading the DCW cluster in other species were aligned using `Clustal Omega v1.2.1` [179] in order to assess sequential conservation. The alignment scores of the resulting multiple sequence alignment were used to construct a phylogenetic tree via neighbor joining in `ClustalX v2.1` [178] with 10,000 bootstraps.

2.1.2 Secondary Structure and Folding Landscape

The secondary structure prediction of *UpsM* was evaluated *in silico* using `RNAfold` [184] without and with constraints on the 3'-flank of the terminator hairpin to not base-pair with the rest of the molecule. Constraining the terminator hairpin from forming, the folding algorithm is able to determine closest alternative conformations, mimicking the transcript during terminator deactivation.

To verify our assumption, we aligned the 5'-UTR regions of all Rhodobacteraceae in our dataset (cropped at the predicted terminator) using `locarna v1.7.16` [181] which respects sequence and secondary structure at the same time. The alignment was folded using `RNAalifold` from the `Vienna RNA 2.0` package [184] with no constraints and, analogously to above, constrained such that the terminator would not be allowed to form.

We further investigated whether *UpsM* could potentially function as a riboswitch by calculating all suboptimal folding structures for the processed sequence of *RSs0682* using `RNAsubopt v2.1.9` and explored the resulting energy landscape using `barriers v1.6.0` to create a barrier tree [184]. Resulting structures were then analyzed for estimated population density over time using `treekin v0.4` [185] and a dot plot of pairing probabilities was produced with `RNAfold v2.1.9` [184] to visually inspect different folding possibilities.

2.2 6S RNA

Degradation of 6S RNA by multiple RNases was investigated based on transcriptomic data and a visualization method developed to yield position-specific transcript borders.

2.2.1 RNA-Seq

RNA-Seq data of different exonuclease knock-out strains was mapped using `segemehl v0.2.0` [151] to the RefSeq genome `NC_000964.3` of *Bacillus subtilis* 168 with the following parameters for paired-end reads: minimal E-value of hits ≤ 0.1, accuracy of matches $\geq 90\%$, minimal match length ≥ 12 and reporting only best matches. Reads in the region of the two 6S RNAs of *B. subtilis* were extracted with a custom script and coordinates set relative to 6S RNA annotation start.

2.2.2 Visualization of RNase Processing

In order to analyze the processing influence of RNases in different knock-out strains, I developed a custom software to visualize the positions of read-pair starts and ends. The tool is called `pe2svg` because it takes `SAM` formated alignments of paired-end short reads and outputs `SVG` formated vector plots. It is written in `Python` and operates in two distinctive modes to visualize read starts and ends in complementary ways. The first (`--mode insert`) shows the alignment of all reads from the first to the last mapped position of each read pair stacked on top of each other sorted by starting position. Reads from different strands can either be merged (`--strandless`) or split into opposing y-coordinate directions. Transcripts with the same start and end position are clustered and displayed as proportional rectangles to reduce image complexity. The second (`--mode bounds`) represents transcript 5'-ends with bars going up and 3'-ends with bars extending down for each position in a coordinate system. Bounding bars can be displayed in absolute value or in percent. Different coloring schemata (gradients `--color 1/2` or predefined color ranges `--color 3/4`) can be used to distinguish the mapping position of the 5'-end. Color ranges were chosen manually to reflect major transcript onsets in 6S-1 ($< 15 < 55 < 110 \leq$) and 6S-2 ($< 20 < 30 < 80 \leq$) relative to the annotated TSS (position `1`). A threshold of cluster size can be assigned to suppress scarcely represented processing artifacts.

For the analysis of 6S-1 and 6S-2 RNA in RNase knock-out strains of *B. subtilis* the alignments were filtered and offset for the two individual loci using `samtools v1.3.1` [152] and a `Python` script (`samoffset`) that normalizes the alignment to given position and strand (offsetting `-o` and reversing `-r` mapped reads if needed). In this case only mapped reads spanning positions 2,814,440–2,814,702 for 6S-1 RNA and 2,095,887–2,096,123 for 6S-2 RNA were extracted and reversed:

```
samtools view -h algn.bam 'NC_000964.3:2814440-2814702' | \
    samoffset -o 2814702 -r > 6S1.sam;
samtools view -h algn.bam 'NC_000964.3:2095887-2096123' | \
    samoffset -o 2096123 -r > 6S2.sam
```

These alignments were then visualized in the two different modes:

```
pe2svg --width 1500 --depth 200 --strandless \
    --color 3 --mode insert --cluster 1 --maxpos 220 \
    --maxval 800000 < 6S1.sam > 6S1.reads.svg
pe2svg --width 1500 --depth 100 --strandless \
    --color 3 --mode bounds --cluster 1 --maxpos 220 \
    --maxval 20 < 6S1.sam > 6S1.ends.svg
pe2svg --width 1500 --depth 200 --strandless \
    --color 4 --mode insert --cluster 1 --maxpos 220 \
    --maxval 150000 < 6S2.sam > 6S2.reads.svg
pe2svg --width 1500 --depth 100 --strandless \
    --color 4 --mode bounds --cluster 1 --maxpos 220 \
    --maxval 20 < 6S2.sam > 6S2.ends.svg
```

2.3 RNase P

The evolutionary development of RNase P RNA and its organellar protein-only counterpart in eukaryotes was analyzed based on bioinformatic structure and sequence predictions, respectively.

2.3.1 Structural Identification of P RNAs

In order to assess the evolutionary landscape of RNase P for the conservation of P RNA we scanned a corpus of 109 completely and partially sequenced genomes from various sources that represent the major branches of the eukaryotic tree (see original manuscript and supplements for details [198]) for structural homologs. We used `Infernal v1.1.1` [180] with `RFAM v12.0` [182] models `RF00009` and `RF01577` which characterize the consensus structure of the canonical nuclear RNase P and RNase P from *Plasmodium*, which represents a deviant structure, respectively. Candidates were additionally screened with `Bcheck v0.6` [199] for conserved regions CR-I–V and the more complex tertiary interaction in paired region P4. Marcus Lechner gathered the aforementioned data and implemented the automatic screening pipeline for P RNAs that provided novel RNase P candidates.

These putative RNase P RNAs were then individually inspected visually, by reconstructing all canonical sub-structures using `RNAfold` [184] and `mfold` [186] to validate overall resemblance with the published minimal consensus RNase P RNA structure [49] including paired regions P1–12 and conserved sequences CR-I–V [50] shown in Figure 5. Structures could not be predicted wholly by existing folding algorithms because of exceeding sequence length and complexity of the long reaching P4 pseudoknot and additional sequence constrains in CR-I–V. Only through careful alignment hairpin by hairpin, we were able to confirm complete or partial conservation of the RNA's hallmarks.

2.3.2 Phylogeny of Organellar PRORP in Eukarya

Similarly to the previous approach, available genomic sequences in databases of NCBI, Ensembl, Bogas, Phytozome, JGI, and Broad using `BLAST` [138] and aligned using `MUSCLE` [200] to verify consensus motifs. This step was carried out by M. Lechner.

A phylogenetic analysis of 88 exemplary selected PRORP sequences representative of all eukaryote groups was performed. PRORP sequences were realigned using `Clustal Omega v1.2.1` [179]. Alignment columns with gaps present in at least 25 % of all sequences were removed to minimize the impact of low-population insertions on the entire alignment. The resulting 384 amino acids long alignment was used in rapid bootstrapping and subsequent search for the best-scoring maximum likelihood tree using `RaxML v8.1.20` [191] with 100 rapid bootstrap inferences under the LG model using fixed base frequencies optimization of substitution rates and the GAMMA model of rate heterogeneity. Optimal amino acid replacement models based on the aligned sequences were determined using `prottest v3.4.2` [194]. The graphical representation of the phylogenetic tree was produced with `SplitsTree4` [201] and carefully optimized for improved readability without changing relational distances.

2.4 circRNA

We identified circRNAs transcriptomic data of honeybee brains and characterized the detected circularized exons using their genomic context as well as DNA methylation data.

2.4.1 RNA-seq of circRNA Enriched Libraries

Another study previously employed a similar strategy successfully for *Drosophila*, where they could show that circRNAs are enriched in data which was not selected for poly(A) tails [68]. We gathered publicly available RNA sequencing data of honey bees from SRA (NCBI Sequence Read Archive ncbi.nlm.nih.gov/sra, accessed 6/7/2018). This non-enriched data was screened individually using the analysis pipeline laid out below. All datasets we could find, however, were selected for poly(A) tails. Samples and projects we used are listed in Table 2.

Table 2: Public RNA-Sequencing libraries used for circRNA screening. Multiple libries of worker bees were downloaded from SRA. Indications are given, whether single-end (SE) or paired-end (PE) sequencing was used.

BioProject-ID	SRA-ID	Caste	Mbases	SE/PE
PRJNA79773	SRR071803	Nurse	449	SE
PRJNA79773	SRR071804	Nurse	495	SE
PRJNA79773	SRR071805	Nurse	944	SE
PRJNA79773	SRR071806	Nurse	1,055	SE
PRJNA79773	SRR071807	Nurse	926	SE
PRJNA79773	SRR071808	Nurse	808	SE
PRJNA79773	SRR071809	Nurse	803	SE
PRJNA79773	SRR071810	Nurse	746	SE
PRJNA79773	SRR071811	Nurse	637	SE
PRJNA79773	SRR071812	Nurse	537	SE
PRJNA79773	SRR071813	Nurse	554	SE
PRJNA79773	SRR071814	Forager	748	SE
PRJNA79773	SRR071815	Forager	752	SE
PRJNA79773	SRR071816	Forager	751	SE
PRJNA79773	SRR071817	Forager	648	SE
PRJNA79773	SRR071818	Forager	642	SE
PRJNA79773	SRR071819	Forager	684	SE
PRJNA79773	SRR071820	Forager	835	SE
PRJNA79773	SRR071821	Forager	755	SE
PRJNA79773	SRR071822	Forager	835	SE

BioProject-ID	SRA-ID	Caste	Mbases	SE/PE
PRJNA79773	SRR071823	Forager	985	SE
PRJNA79773	SRR071824	Forager	871	SE
PRJNA79773	SRR071825	Forager	867	SE
PRJNA200755	SRR838836	Nurse	278	SE
PRJNA200755	SRR838837	Forager	291	SE
PRJNA200755	SRR838838	Nurse	282	SE
PRJNA261549	SRR1582004	Nurse	19,563	PE
PRJNA261549	SRR1582203	Forager	12,955	PE
PRJNA104931	SRR446005	Forager	34,472	PE
PRJNA104931	SRR446006	Forager	37,189	PE
PRJNA104931	SRR445999	Nurse	41,193	PE
PRJNA104931	SRR446000	Nurse	56,909	PE
PRJNA104931	SRR446001	Nurse	29,274	PE
PRJNA104931	SRR446002	Nurse	29,759	PE
PRJNA104931	SRR446003	Nurse	26,707	PE
PRJNA104931	SRR446004	Nurse	43,535	PE
PRJNA104931	SRR446007	Forager	23,339	PE
PRJNA104931	SRR446008	Forager	31,705	PE
PRJNA104931	SRR446009	Forager	34,610	PE
PRJNA104931	SRR446010	Forager	26,855	PE

For our own libraries, bees were derived from colonies with normal age structure and with a naturally mated queen located on the grounds of the University of Würzburg. Individuals were considered as nurse bees, if they clearly poked their head into open brood cells containing young larvae. Foragers were captured when returning from a foraging flight and having huge pollen loads at their hind legs. Collected bees were frozen in liquid nitrogen immediately. A single cohort colony was established by transferring 2,500 newly emerged bees (marked by the same color immediately after hatching) into a small hive together with one queen in one brood frame and one frame with pollen and honey. Single cohort colony bees were collected at the age of eleven days and controlled for their social task.

We used a total of four RNA-Seq libraries to determine circular transcripts present in the brain of honeybees. First, an enrichment control was compiled from the brains of ten dissected nurse bees and ten dissected foragers. Total RNA was extracted with Isol-RNA lysis reagent (5PRIME, Hilden, Germany) and treated with DNase I. The sample was divided into two halves. One half (E_+) was treated with 3 units RNase R (epicentre, Madison, USA) per μg of total RNA to deplete linear transcripts. Digestion was performed for 30 min at 37 °C. For the other half (E_-) an equivalent volume of double distilled water was added. Afterwards, both samples were purified using phenol–chloroform extraction. Efficacy of the RNase R treatment was verified in a control experiment shown in the supplements of the manuscript. Second, we took additional samples from ten nurses and ten foragers separately and treated both with RNase R as described above (samples F_+ and N_+ , respectively) in order to distinguish task dependent expression levels. Samples were collected and prepared by Christoph Erbacher and Markus Thamm at the University of Würzburg. Library preparation and Illumina© sequencing (125 nt paired-end) were performed by GATC Biotech AG (Konstanz, Germany). All RNA-Sequencing data was made publicly available via bioproject `PRJNA345404` and are listed with further details in Table 3.

Table 3: circRNA specific RNA-Seq libraries.

SRA ID	Name	Sample description	Reads
SRR4343845	E_-	background no RNase R treatment	8,432,479
SRR4343846	E_+	background with RNase R treatment	7,690,777
SRR4343847	N_+	nurse bees with RNase R treatment	5,843,829
SRR4343848	F_+	forager bees with RNase R treatment	5,931,097

Identification of circRNAs

We used two independent algorithmic approaches for the identification of circular RNAs. In one approach reads were mapped to the NCBI *A. mellifera* genome `v4.5 rel102` (RefSeq `GCF_000002195.4`) using `segemehl v0.2.0` with the split reads option (`-S`) [153]. The alignment was subsequently screened for model-free splicing events using the accompanied `testrealign` tool. In the second approach we used `BWA v0.7.5a` [148] as mapping tool and subsequently screened the alignment with `CIRI2 v2.0.6` using default parameters [167]. Identified junctions were post-processed using custom scripts bundled in our `Chiasm` suite. `Chiasm` was also used to perform the statistical calculations later on (e.g. CpG-content, pairing-probability, see below). The full analysis pipeline is publicly available at git.io/chiasm. More precisely, junctions with almost identical start and end positions were merged if they differed by less than 6 nt. Junctions mapped ± 5 nt next to exon boundaries were corrected to exactly match the boundary. This accounts for small variations in sequencing and mapping, e.g. due to flanking intron sequence being potentially identical to the junctioning exon or indels in the genome [162]. Only junctions with a total of ten or more JSRs were considered for screening in public data and our own experiments.

BSJs were annotated by assigning them to overlapping gene regions first and determining the longest possible transcript containing the same exon boundaries (where matching exactly). In parallel, the total read number for each gene (including all isoforms) was counted using `featureCounts v1.5.1` [171]. Analogously to present studies in *Drosophila* [68] we normalized BSJ read counts, $norm(n_\circ)$, by dividing the number or circular JSRs, $n_\circ$, by mapped library read count, N, in million, divided by reads per kilobase million (RPKM) of the host gene, g. The latter is defined as number of reads assigned to the host gene, n_g, divided by the length of the gene, l_g, in megabases and divided by library size of mapped reads, N in Mb.

$$norm(n_\circ) = \frac{n_\circ}{\frac{N}{1{,}000{,}000} \cdot RPKM_g} \tag{2}$$

with

$$RPKM_g = \frac{n_g}{\frac{l_g}{1{,}000} \cdot \frac{N}{1{,}000{,}000}} \tag{3}$$

which can be simplified to

$$\Rightarrow norm(n_\circ) = \frac{n_\circ \cdot \frac{l_g}{1{,}000}}{n_g} \tag{4}$$

Identified circRNAs were divided into two sets of different stringency levels. The low-stringency set contains all circRNAs picked up by both approaches (`testrealign` and

CIRI2) with at least three JSRs. In the high stringency set, we only considered BSJs with more than ten JSRs across all libraries as suggested in literature [68]. Thereby, the BSJ has to be found in library E_+ and at least one other independent RNase R treated library. Following the recommendation in [163], potential circRNAs had to be at least 5-fold enriched in E_+ over the E_- control to ensure their circularity.

Table 4: PCR primers. Convergent (L) and divergent (C) primer pairs consisting of forward (+) and reverse (-) sequences for each amplified locus. Divergent primers span the BSJ with the stated product length in nucleotides.

circRNA (locus)	Form	Product length	Sense	Sequence 5'–3'
ame_circ_0001970 (LOC413427)	linear	114 nt	+	GAGGTGGAAAACGGGGAACC
			-	GCTGATGTCAATTCTCGGCG
	circular	108 nt	+	ACGCCGAGAATTGACATCAGC
			-	AGCCCTCGATGATGCTGCG
ame_circ_0002142 (LOC410393)	linear	125 nt	+	TACGAGGGTGCGACACG
			-	CTCATCCCCATTCCGTCGC
	circular	111 nt	+	GGATCCTCTCGGGTCAACG
			-	GCACCCTCGTACCATGCC
ame_circ_0000163 (LOC408576)	linear	117 nt	+	CTACAGCCACTTCCGGTCC
			-	GTCTCGGTGATCGACTCAAGC
	circular	91 nt	+	ACCGTCCTCCTCTTCTACTCG
			-	ATGTCACAGGTGTGGGAACC
ame_circ_0002577 (LOC409655)	linear	115 nt	+	CCCCGCAAAGATTGAAACCG
			-	TCGAGTTCACTCAACGGACC
	circular	99 nt	+	CATGGCCTTCGTAAACCAGC
			-	AGAGCGTGTGACATGTACGG
ame_circ_0002579 (LOC409655)	linear	114 nt	+	ATGCACTCTTACGAAACGCC
			-	CAGTGGTGGTGTTTTCCTGC
	circular	89 nt	+	CGGATCCCATCCACAGATTCG
			-	TTAAATCACCCTCTCTCACCCC
ame_circ_0000721 (LOC724885)	linear	114 nt	+	ACTACCCTTACAACGTGTCCG
			-	TCTTCTTTCGGGGGTGTTGC
	circular	118 nt	+	GGCATAGTGCCCGACTACC
			-	TATTCGCTCGCTCCAACCG
ame_circ_0001286 (LOC411534)	linear	94 nt	+	CATGGCGGAGAAACAACGC
			-	CATCTTCCTCGGGCTTTCGG
	circular	84 nt	+	CAGTCGGCAAGTTCAAGAGC
			-	TGTCTCGGCGTTGTTTCTCC
ame_circ_0001822 (Rsmep2)	linear	298 nt	+	GGTTGAAAGTGGTCGGCG
			-	GATGGCGGTGAGGAAGACC
	circular	191 nt	+	GATGGCGGTGAGGAAGACC
			-	GGTTGAAAGTGGTCGGCG
ame_circ_0001852 (CoRest)	linear	121 nt	+	CGTGTCCTTTTGGCGTCG
			-	CGTCAGTGCCTTGGCTTC
	circular	116 nt	+	CGTCAGTGCCTTGGCTTC
			-	CGTGTCCTTTTGGCGTCG
ame_circ_0000398 (LOC724851)	linear	226 nt	+	GCCGAGTAAAAATGGCGC
			-	GGGAATCCGGGACATTGC
	circular	186 nt	+	GGGAATCCGGGACATTGC
			-	GCCGAGTAAAAATGGCGC
ame_circ_0000232 (Mup2)	linear	363 nt	+	CGAAACAAGTCACGCGG
			-	CCTCTTCTATTGGCGACTACG
	circular	379 nt	+	CCTCTTCTATTGGCGACTACG
			-	CGAAACAAGTCACGCGG
ame_circ_0001780 (rad)	linear	1835 nt	+	CTGGTGTGGCAAGGTGG
			-	TGGATGGGACGATCTTATGC
	circular	1832 nt	+	CCTCTTCTATTGGCGACTACG
			-	CGAAACAAGTCACGCGG
EF1α	linear	118 nt	+	ACGCTATATTACCGCCGTCC
			-	AAAATACCGGTCTCCACCCG

Validation of Selected circRNAs Through PCR Amplification

Total RNA was extracted from ten worker bee brains and prepared as described for the RNA-Seq preparation above (also with and without RNase R treatment). After DNA digestion, 1 μg of RNA were transcribed into cDNA using RevertAid H minus reverse transcriptase (Thermo Fisher Scientific) adhering to the manufacturer's specifications. For PCR amplification 15 μmol of divergent or convergent primers were added to 10 ng of cDNA with 25 μL of Phusion Polymerase master mix (Thermo Fisher Scientific). PCR steps consisted of 30 sec heating to 98 °C followed by 35 cycles of 10 sec denaturation at 98 °C, 10 sec annealing at 62 °C and 8 sec elongation at 72 °C. After a final extension period of 10 min at 72 °C, PCR products were either stored at -20 °C or subjected to agarose gel electrophoresis prestained with 5 µL of GelRed (Biotium). Primers used for PCR amplification are listed in Table 4.

Further PCRs with divergent primers and qPCR experiments with TaqMan probes for enrichment control and quantification circRNAs were conducted by Christoph Erbacher and Markus Thamm at the University of Würzburg. Those methods are outlined in the attached manuscript.

2.4.2 Characterization of Candidate circRNAs

We extracted whether the circRNA contained part of the 5'-UTR, 3'-UTR of a canonical protein-coding transcript or if it exclusively contained coding regions. The number of exons of the longest fitting transcript between the 5'- and 3'-end of each BSJ was noted along with the index of these exons within the transcript. For exonic sequence, all exons between the junction sites in the transcript were considered. Flanking introns were determined by including the sequence outside of the BSJ exon boundary until the next exon in the same transcript. Splice signals were visualized using `WebLogo v3.4` [202].

Statistical Control for Circularized Exons

For a statistical comparison of the annotation properties between circularized exons and non-circular transcripts a random control set was generated. Annotated internal exon boundaries within transcripts consisting of more than two exons were randomly drawn from all chromosomes weighted by their lengths. Genes harboring any JSRs found in this study were excluded from this control. 10,000 such BSJ were generated and all analysis steps after quantification of circRNAs identified in actual data replicated with this list, to provide a statistical control for intron-exon structure analysis.

Complementarity of Intron Sequences

In order to screen for complementarity between flanking intron pairs, the full length 5'-intron was matched to the 3'-intron using `BLAST v2.2.28+` [138] with a word size of six to determine the highest scoring stretch of reverse complementarity. We repeated the procedure with 100 nt from the end of the upstream and 100 nt from the start of the downstream intron, to discern whether especially approximate regions showed increased complementarity. The same 100 nt portions were used for structural analysis

utilizing RNAcofold [184]. We applied soft constrains to ensure MFE scores solely based on base-pairing between both intronic regions. Both procedures were repeated with all combinations of starts and ends of the respective introns as an educated control set (an interaction of the end of the upstream and the end of the downstream intron is probably not relevant). Surprisingly, the results for all combinations were similar. To rule out, that we bias for specific length effects at 100 nt, all calculations were also done with 50 and 200 nt without changing the outcome. Introns were checked for GC-content ignoring undetermined residues in the genome sequence (N). Similarly the mononucleotide frequency of cytosine and the relative frequency of CpG dinucleotides was calculated.

Homology Screen and Functional Annotation

Predicted circRNAs were correlated to those previously reported for *D. melanogaster* and *B. mori* [68,73,87]. We matched the loci based on the predicted homologs of the closest protein-coding gene with respect to OrthoDB v9 [203]. CircRNAs from genes without homolog could thus not be accounted for. Homologous fruit fly genes were then submitted to the online PANTHER annotation platform for further over-representation analysis using Fisher's Exact test with FDR for multiple testing correction. We included functional annotations with more than 5-fold over-representation and FDR below 1 %.

2.4.3 DNA Methylation

To assess whether the observed increase of potential DNA-methylation sites is reflected in actual DNA-methylation, we use whole genome bisulfite sequencing data of worker bees that was publicly available. Precisely, we used all native worker libraries provided BioProject PRJNA104931 [92] and combined them for this analysis (Table 5).

Table 5: Bisulfite sequencing libraries of honeybee worker brains. Sequencing data from another study [92] was analyzed to determine the methylation state of specific regions around identified circRNAs and control exons.

SRA-ID	name	caste	Mbases
SRR445809	W5	worker	8,183
SRR445808	W4	worker	7,152
SRR445807	W3	worker	10,559
SRR445806	W2	worker	6,553
SRR445805	W1	worker	5,602
SRR445799	N1	nurse	24,139
SRR445778	N2	nurse	14,082
SRR445777	N3	nurse	15,342
SRR445776	N4	nurse	24,317
SRR445775	N5	nurse	26,406
SRR445774	N6	nurse	6,079
SRR445773	F1	forager	15,639
SRR445771	F2	forager	17,026
SRR445770	F3	forager	15,720
SRR445769	F4	forager	7,377
SRR445768	F5	forager	12,286
SRR445767	F6	forager	8,953

Reads from whole genome sodium bisulfite sequencing libraries were mapped to the honeybee genome with reduced nucleotide alphabet (`C->T` and `G->A` converted) with `Bowtie v2.2.6` [154] using the recommended parameters [169]. Methylation patterns were analyzed using `Bismark v0.19.1` [169] also with default parameters suggested by its authors. The methylation status of each covered cytosine was aggregated throughout the different libraries with a custom script to determine the average methylation per base (strand unspecific) for each intron individually. An average coverage of at least five reads was required for each intron in order to reliably determine the methylation status. Calculations were done for 50, 100 and 200 nt as well as for the length of the complete intron where it exceeded 200 nt and normalized by the respective sequence length. This way absolute methylation is compared instead of relative methylation, because we previously established that CpGs are more common in circRNA flanking introns. A single-sided Wilcoxon-Mann-Whitney rank-sum test [204] was used to determine significance of the methylation increase over the control. Additionally, a position-specific methylation profile was generated relative to the splice site by accumulating mC/C ratios over all experiments and dividing by the number of intron sequences spanning the covered position. No significant differences in average methylation was found between nurse and forager bee libraries for the genes relevant in this study.

2.4.4 miRNA Interference

Predicted and experimentally verified miRNA sequences of *A. mellifera* were obtained from `miRBase rel21` [205]. Potential target sites were screened in all exon sequences overlapping with the identified circRNAs using nucleotide two to seven of the mature miRNA sequence, see [123]. We implemented a miRNA target prediction based on extended regular expression matching that is published along with the sources of the identification pipeline. The algorithm finds matches that are reverse complementary seed region of 6 nt starting with the second base of the mature miRNA sequence and extends them to at least 15 nt non-consecutive base-pairs with at most one mismatch. In direct comparison, our program yields similar results to `miranda` [206] but provides additional information about the interaction duplex to better assess the validity of the hit.

For each potential miRNA binding site, we determined conservation in further *Apis* species (*A. cerana*, *A. dorsata*, *A. florea*) and other eusocial insects (*E. dilemma*, *L. ventralis*, *M. quadrifasciata*, *B. impatiens*, *B. terrestris*) for the seed region with 100 nt up- and downstream using the best `BLAST` match [138] in the respective genome. We considered a site conserved if the 6 nt seed region was perfectly conserved among three out of four *Apis* or four out of five eusocial insects, respectively.

As random control we used linear exons, see Section ‘Statistical Control’. We split the control to sets of about equal size (42 sets) and applied the above procedure to each set. This results in 42 control datasets where each represents a subset of exons with similar length to avoid a bias due to an over-representation of certain length species. Identified target sites were normalized to sites per 1,000 nt.

3 Results and Discussion

In the following, individual findings that were immediate results of the bioinformatical contributions to the four studies presented in this thesis are highlighted. These summarized excerpts are augmented with complementary results, which are not part of the main articles. Details of all the results can be found as part of the respective manuscripts attached in Section 5. For each part the implications of the transcriptomic and genomic analysis is discussed in the context of the respective ncRNA and its transcriptional processing step.

3.1 Characterization of UpsM

Like other gram negative and rod shaped bacteria, *R. sphaeroides* possesses a syntenically conserved DCW cluster that aides in the well timed cell devision process. However, our experimental transcriptomic data treated with TEX shows an unusual elongation of the 5'-UTR 268 nt upstream of *mraZ*, the first gene in the operon. A distant upstream promoter gives rise to expression of an abundant ncRNA which had been generically classified as an orphan sRNA in a previous small RNA screening [23]. Describing its unique locus the 206 nt long ncRNA was henceforth renamed to **up**stream **s**RNA of **m**raZ (UpsM). We predicted a strong intrinsic terminator 84 nt upstream of *mraZ* through *in silico*. Indeed, most of the transcripts in the RNA-Seq libraries end at this particular feature while only a fraction of transcripts extend into the first gene of the cluster. This observation was corroborated by rapid amplification of cDNA ends and qPCR. Because *upsM* shares its promoter with the rest of the gene cluster, it is extremely likely that this ncRNA bares at least an indirect role in the regulation of the rest of the operon.

We further found a potential ORF in the same UTR which would encode for a 56 amino acid long peptide. However the transcription terminator is situated directly between start and stop codon, thus rendering the transcribed UpsM the ORF independent features. Structural prediction of the encoded peptide revealed no resemblance to known protein domains and mass spectrometry of whole cell lysate did not detect peptides of the predicted mass or sequence. It is thus unlikely that this ORF is translated *in vivo*.

3.1.1 Conserved but Unique to Rhodobacteraceae

In order to shed light on a potentially conserved functional role of *UpsM* further publicly available RNA-Seq data from related species with *mraZ* as the first gene in the DCW cluster was analyzed.

Figure 8 shows similar expression patterns in the 5'-UTR in *R. capsulatus*, *R. pomeroyi*, *P. gallaeciensis* and *D. shibae* where transcripts start approximately 150 nt upstream of *mraZ* but reads diminish after 70–200 nt at similarly plausible predicted intrinsic terminators. Throughout the considered Rhodobacteraceae, transcripts similar to UpsM are expressed in the absence of (or at least in many fold higher amounts than) *mraZ* expression. In contrast, other species apart from Rhodobacteraceae typically have an exclusive TSS closer to *mraZ* resulting in 5'-URTs of ≤ 70 nt, including the closely related Alphaproteobacterium *C. crescentus*. Moreover, upstream terminators are either not found at all or predicted to be rather weak in those species. An ncRNA annotated in the 5'-UTR upstream of *mraZ* in *S. meliloti* does not match the transcripts detected in the sequencing data. Nevertheless, caution must be exercised during the interpretation of divers datasets from different studies that were conducted under other experimental conditions. Additionally, the phylogenetic tree (left side of Figure 8) based on the 300 nt upstream of *mraZ* clusters the species according to their class within their Proteobacteria phylum which suggests that this region also stands under some evolutionary pressure.

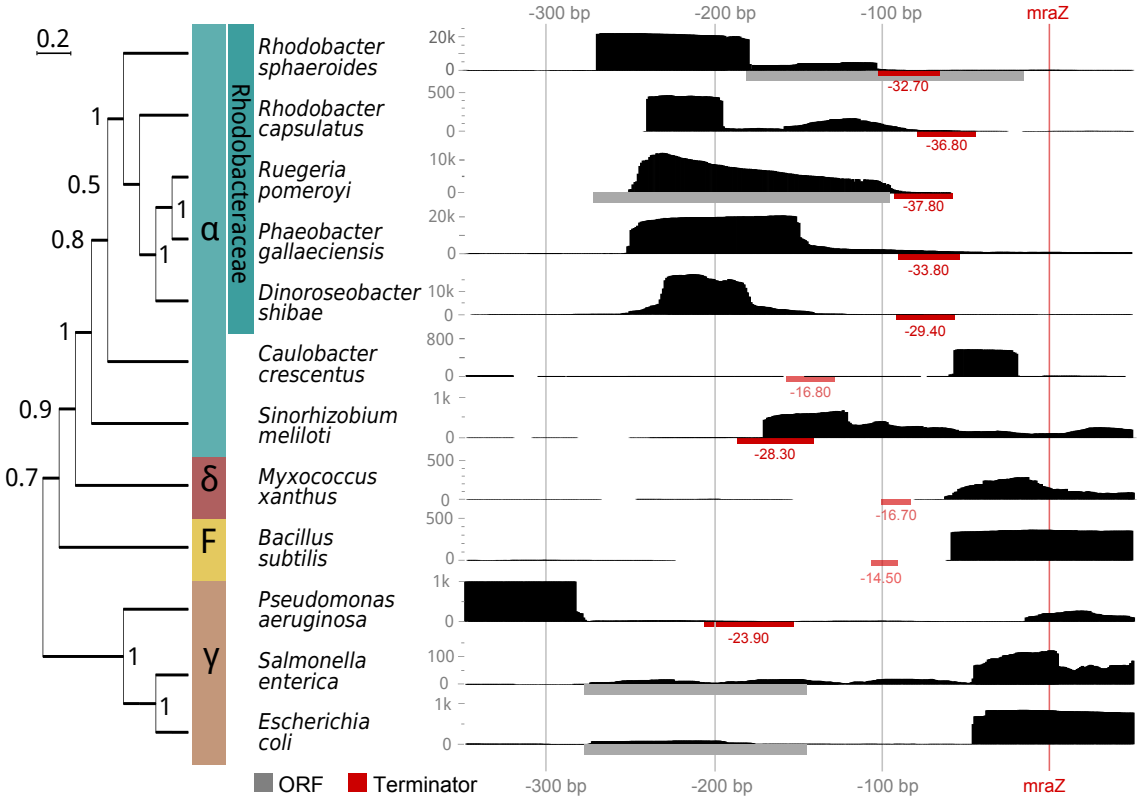


Figure 8: Transcription analysis upstream of the DCW cluster in different species. Read coverage of the 300 nt upstream of *mraZ* (red line) in libraries of each species is presented as black bar charts (right). Sequencing depth varies as indicated by each Y-axis. The strongest terminator (red) including the its MFE in kcal/mol as determined by `RNAfold` is indicated below each graph along with potential ORFs (gray). The phylogenetic tree with bootstrap values based on alignments of the entire 300 nt cluster the different classes of Proteobacteria.

Taken together our data suggests that a long *mraZ* 5'-UTR with intrinsic terminator generating an sRNA combined with no separate TSS for *mraZ* is an exclusive feature of the family of Rhodobacteraceae.

3.1.2 Potential Riboswitch Characteristics

Due to the terminator hairpin at the 3'-end of UpsM, the transcript might regulate transcription read-through by structural rearrangement. A consensus structure of the identified UpsM homologs revealed four structured regions: R1 is hardly conserved because the transcripts in other Rhodobacteraceae is shorter at the the 5'-end. R2 forms a long hairpin with a small bulge while R3 consists only of 4 immediate base-pairs. R4 forms the terminator hairpin at the 3'-end of the ncRNA. Figure 9A shows the predicted MFE for UpsM in *R. sphaeroides*. The secondary structure elements R2/3 and R4 show similarities to typical riboswitches (R2/3 aptamer region, R4 terminator) [5].

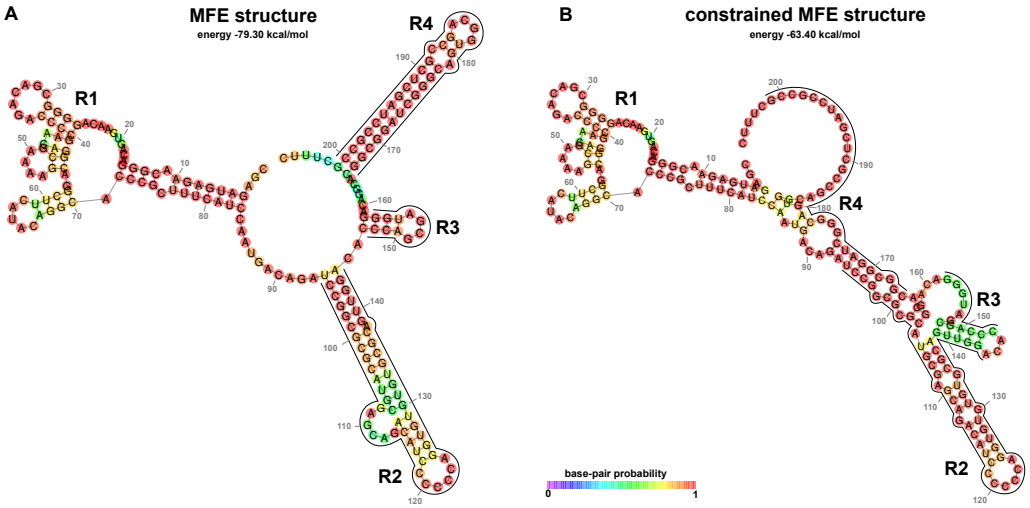


Figure 9: Structural analysis of upsM. Analogous structured regions are indicated as R1-R4. RNAfold structure of upsM in *R. sphaeroides* with (A) and without (B) constraint terminator (R4).

Constraining R4 from forming resulted in alternative base-pairing with R2 in *R. spaeroides* which resulted in a net loss of 15.9 kcal/mol in free energy, shown in Figure 9B. A similar reconfiguration was displayed in consensus structure predictions. This shows that generally there is a potential for switching to a different conformation within the ncRNA but a potential factor, i.e., a ligand would need to contribute a considerable amount of free energy to the initial structure. A second *in silico* approach to assess the possibility of a riboswitch in this region by exploring the folding landscape of intermediate transcript lengths (Figure 10) did not reveal suboptimal structures that would break off the terminal hairpin in R4. None of the most probable suboptimal structures of UpsM differ in the terminator hairpin (R4). States 1 and 4 represent the most pronounced folding minima with the highest probability. Potential interactions between the structured regions are not observed. The population density graph of the four most pronounced species in Figure 10B shows no actual competition between alternatively substructures exists, as structure 1 is the predominant species throughout the simulation. These findings indicate a rather stable structure that requires an additional partner (e.g. RNA, protein or ligand) for substantial refolding.

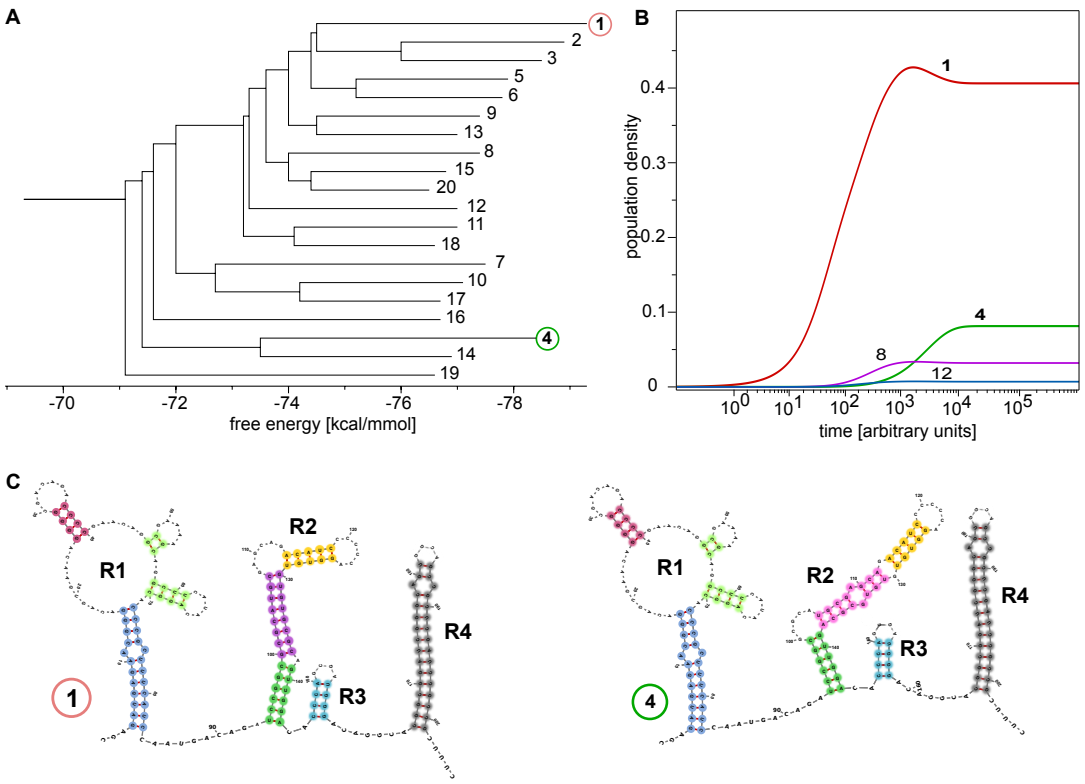


Figure 10: Folding landscape of UpsM. (A) Barrier tree of all suboptimal structures of the processed RNA where four main species with favorable energies were identified (1 & 4). (B) Population density of these structures over time for folding. (C) Structural representations of these related structures colored by similar elements.

This study exemplifies the strength of specifically conducted transcriptomic experiments for the characterization of an ncRNA. However, it also demonstrates the lack in mechanistic explanation of the transcription process and offers a combination of comparative transcriptomics together with structure prediction to offer insights into the regulatory function of UpsM in this case. Further validation of the potential riboswitch functionality has now to be investigated experimentally to clarify under which conditions read-through can occur.

3.2 RNase Degradation of 6S RNA *in vivo*

Paired-end transcriptomic sequencing of multiple RNase knock-out strains in *B. subtilis* is able to resolve degradation sites of the 6S-1 and 6S-2 RNA transcript *in vivo*. The method of processing mapped reads that was developed as part of the analysis allows for a straightforward interpretation of transcript starts and ends. Results are shown for both 6S RNAs in Figure 11 and 12, where the distribution of transcripts sorted by 5'-position in the 6S gene is presented for each of the strains. A second simplified version of read-pair starts (up) and ends (down) is shown below each distribution. The second version particularly facilitates comparison of ribonuclease activity as unique read onsets and falloffs can immediately be spotted. Additionally, transcript 3'-ends in the bar chart are colored according to their 5'-starting position to give an indication of the length of the whole transcript. Colors were chosen to reflect common transcript onsets of both 6S RNAs individually throughout the data (see Methods Section for details). Commonly

used representations of mapped reads, i.e., in IGB [207] or IGV [208] show either only the number of mapped reads per position and thus obscure information about actual read ends, or cannot show all reads for a deeply sequenced locus at once.

3.2.1 6S-1 RNA Maturation by RNase J1

Immediately visible is the different number of transcripts for both 6S RNA variants throughout the libraries. Only $\Delta rnjA$ and Δrny exhibit comparable read levels (~150k) between the different loci. Curiously, relative expression levels between the libraries are switched, as 6S-1 RNA shows about 5x lower read numbers for the RNase J1 and Y knock-outs compared to wild-type, but 6S-2 RNA shows 5x higher read numbers in the same libraries compared to wild-type.

The RNase J1 knock-out exhibits virtually no 6S-1 RNA transcripts starting at position +1 whereas the vast majority of all reads start at this position in the other libraries. Instead 6S-1 RNA contains an extra 11 nt at the 5'-end in 10 % of the transcripts and extends up to 3 nt after the final uracil at position 190 in approximately 40 % of all reads.

3.2.2 6S-2 RNA Starts at Position +10

RNase PH is visibly responsible for 3'-maturation of 6S-2 RNA as only the Δrph and quadruple knockout (which includes Δrph) exhibit read ends beyond the uracil at position +211. Transcripts with an intact 5'-end at position +1 make up the majority in all libraries were both RNase J1 and Y are present. When one of them is knocked out, only ~15–30 % of the reads contain these 9 extra nucleotides. In all libraries, transcripts from position +1 almost never reach the full length of 211 nt as opposed to transcripts starting at position +10 or +44 and beyond. Knock-outs of RNase J1 and Y produce a smaller proportion of short transcripts from position +1 to +60 which suggests that they are responsible for 5'–3'-degradation starting at the end of the central bulge.

The particular challenges of this degradation study exemplifies the need for custom tailored analysis of RNA-Seq data in order to harness the power of its single nucleotide resolution. Commonly available approaches only regard the read coverage for each position and would have completely missed the observations we were able to make. Reducing the data to a visualization of transcript starts and ends presents a much more comprehensive approach to this kind of investigation. The same data and tool can now be used to compare degradation of other loci to characterize the roles of the individual RNases further on one hand and explore global mRNA stability on the other.

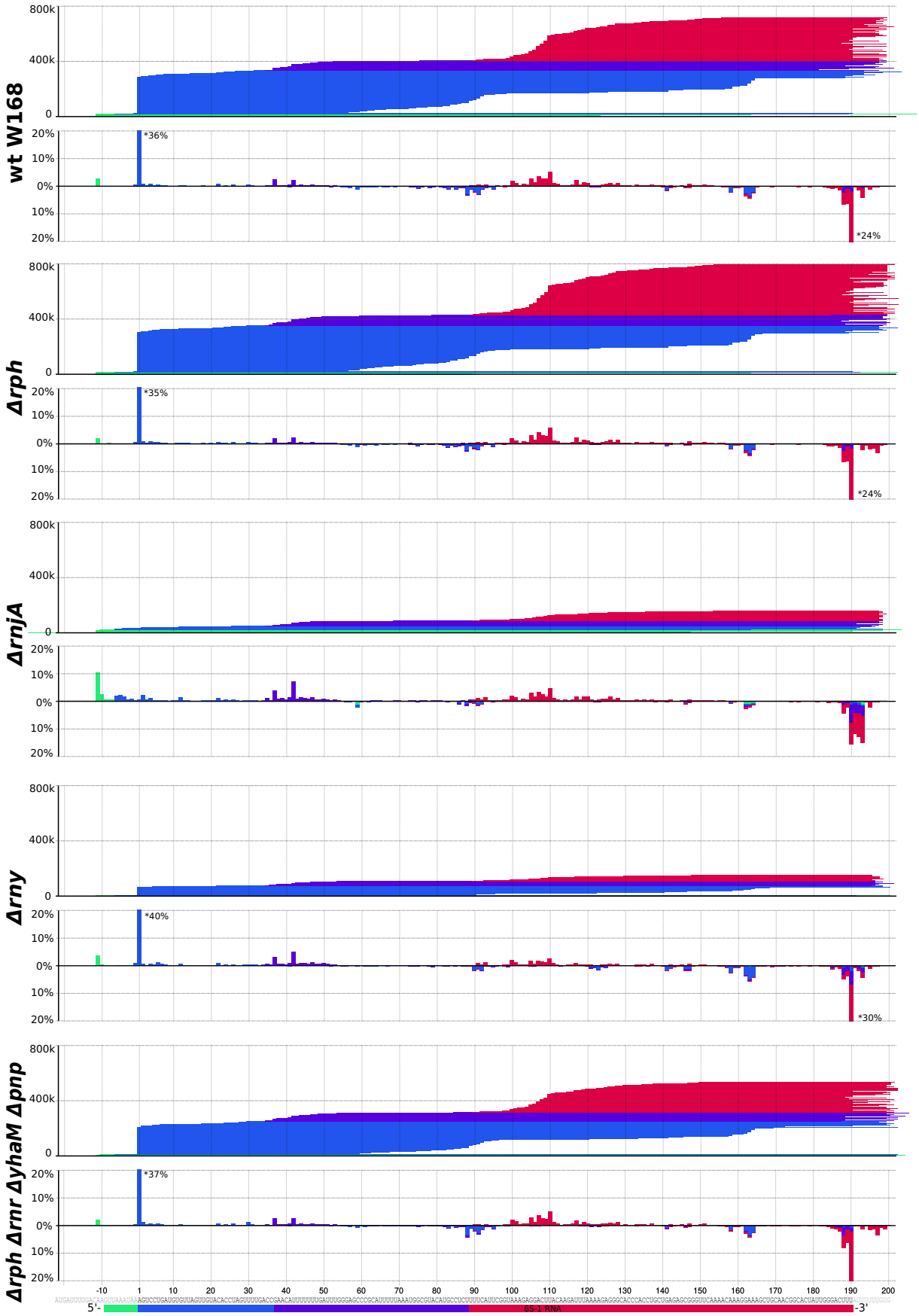


Figure 11: 6S-1 processing in different exonuclease knock-outs. Transcripts of paired-end RNA-Seq libraries of one wild type and four different knockout strains. Each graph shows the distribution of full length transcripts mapped to the locus sorted by starting position relative to the annotated 6S-1 RNA gene (sequence at the bottom). The bar chart below each shows the number of transcript onsets (5'-end, up) and falloffs (3'-end, down) colored by transcript starting position. The Y-axis is the number of mapped reads and percentage within the locus respectively.

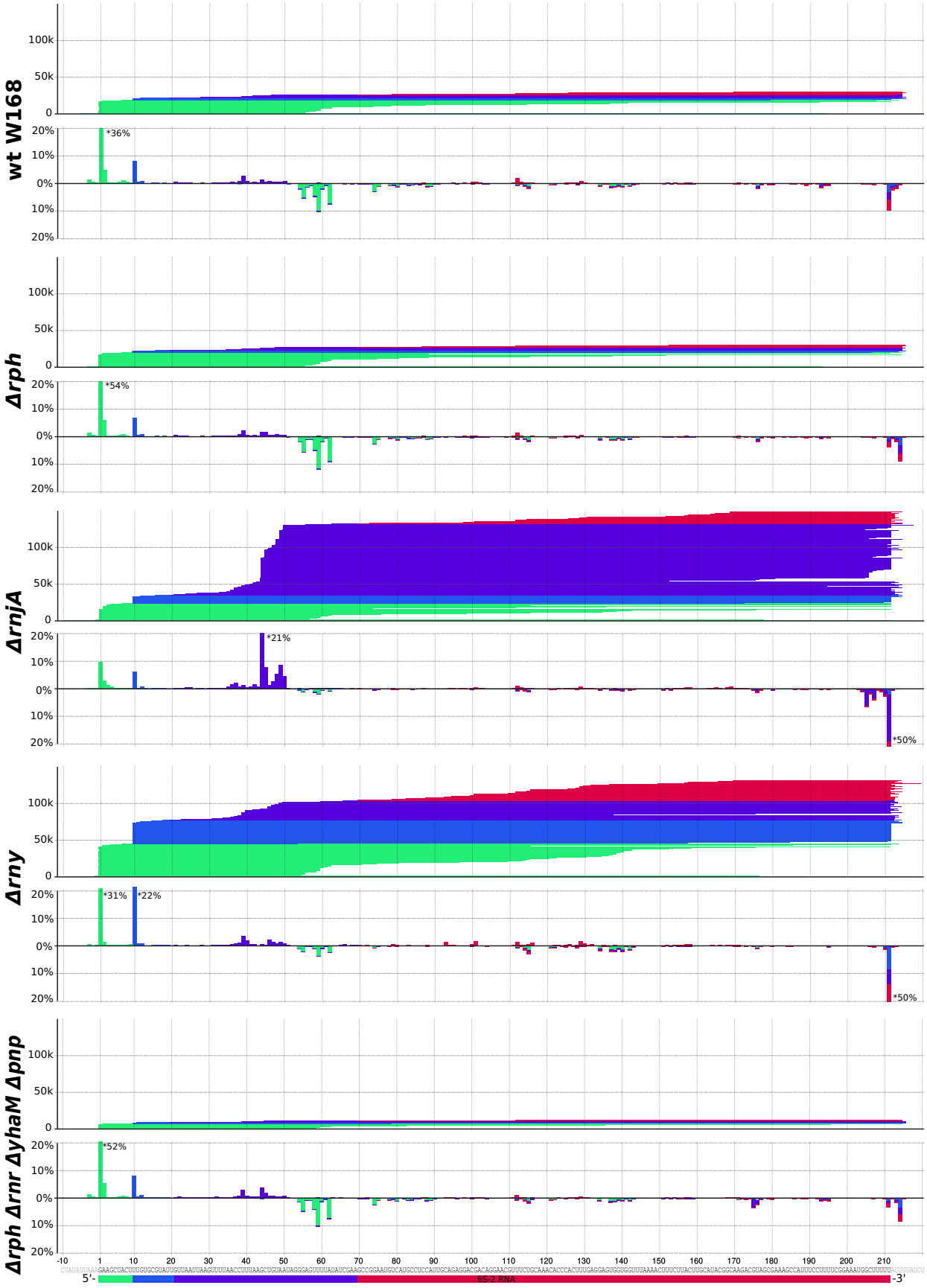


Figure 12: 6S-2 processing in different exonuclease knock-outs. Transcripts of paired-end RNA-Seq libraries of one wild type and four different knockout strains. Each graph shows the distribution of full length transcripts mapped to the locus sorted by starting position relative to the annotated 6S-2 RNA gene (sequence at the bottom). The bar chart below each shows the number of transcript onsets (5'-end, up) and falloffs (3'-end, down) colored by transcript starting position. The Y-axis is the number of mapped reads and percentage within the locus respectively.

3.3 Evolution of Eukaryotic RNase P

In an effort to analyze the occurrence of organellar P RNA and evolution of PRORP based on the increasing number of completely and partially sequenced genome, we identified previously not annotated RNase P RNAs and protein-only variants in a representative selection of eukaryotes.

3.3.1 Diversity of Organellar RNase P RNA

A total of 29 likely genuine organellar P RNAs were identified and structurally characterized in eukaryotic organelles with the help of available sequencing data. Annotating this ubiquitous ribozyme, helps to further understand RNase P catalytic activity in general and reshape existing structural models to reflect possible deviations from the consensus. Especially insertions or striking modifications to the P3 and P8/P9 region were apparent among the newly annotated structures. Table 6 lists all reasonably predicted RNase P RNAs characterized by our approach with respective idiosyncrasies to the consensus described in [49] and [50].

Table 6: RNase P RNAs structurally identified in Eukarya. List of identified P RNAs through structural screening of newly sequenced organisms with conserved pairing regions (P) and complementary tertiary regions (CR) adhering to the minimum consensus structure [49] and signature motifs [50] for organellar P RNA.

Organism	Functional?	Comments
Cyanophora paradoxa	questionable	no 5'-terminal A in CR-II; hairpin between P2 & P3; deviation from CR-V consensus
Porphyridium purpureum	yes	insert between eP5/7 and P8; extended P12, P15 & P19
Galdieria sulphuraria	yes	see *Porphyridium purpureum*
Naegleria gruberi	yes	
Monosiga brevicollis	yes	insertion in P3 & CR-II
Capsaspora owczarzaki	yes	mismatches in P1; insert in P3; extension of P9; bulged nt in P8 and P9
Amoebidium parasiticum	yes	
Acropora digitifera	yes	multiple candidates including possible pseudogenes
Hydra magnipapillata	possibly	no CR-II; identical in *Hydra vulgaris*
Nematostella vectensis	yes	multiple candidates including possible pseudogenes
Mnemiopsis leidyi	likely	weak eP5/7
Pleurobrachia bachei	yes	multiple copies; weak P2; non-canonical bp in P3 & P12
Ascaris suum	yes	deviation from CR-III consensus; weak P2; insert in P8/P9
Brugia malayi	yes	
Caenorhabditis elegans	yes	deviating CR-II signature "AGAAG"; weak P2; similar in *C. briggsae*, *C. brenneri*, *C. japonica* & *C. remanei*
Necator americanus	yes	insertion in P8/P9 domain
Steinernema monticolum	yes	insertion in P8/P9 domain; extended eP5/7
Trichuris trichiura	yes	
Trichoplax adhaerens	yes	
Amphimedon queenslandica	yes	insertion in CR-II
Allomyces macrogynus	yes	two copies with > 300 nt insert between eP5/7 & P10/11; absence of canonical P8 & P9
Gonapodya prolifera	possibly	four copies; insert of two Us in CR-I; deviation from P4 consensus but retained complementarity
Rhizophydiales sp.	no	deviation from CR-I,-II & -V consensus; no P4 helix
Spizellomyces punctatus	no	degenerated P1; potential U bulge in CR-V; small P3 deviation from P4 consensus but retained complementarity
Mucor circinelloides	yes	
Rhizomucor miehei	yes	
Rhizopus oryzae	yes	
Oxytricha trifallax	yes	weak P8 & P9; unusual P3; extended P19
Thecamonas trahens	yes	4nt insertion between P10/11 & eP5/7

Figure 13 shows three example structures from the prediction process. While the predicted structure in *C. elegans* (A) reflects the model presented in Figure 5 despite a relatively short P2 and minor discrepancies in the CR-II signature sequence. Additionally, a short P15 upstream of CR-IV is also predicted to form in *C. japonica*. Closely resembling structures were also found in other *Caenorhabditis* species, which corroborates its functionality. Some predicted structures were significantly larger and more complex. One example is the structure in *G. suphuraria* (B) with an extremely long but stable P12 hairpin, an additional short hairpin before P8 and P9 and bifurcation of an unusually long P19. Despite the structural deviations from the previous model, signature sequences are largely intact and we decided to classify this molecule a functional P RNA. In contras, other structures resembled the model extremely closely, but were rejected due to non-canoncical sequence motifs in CR-I and V which prevent P4 formation, as in the case of *Rhizophydiales* sp. (Figure 13C).

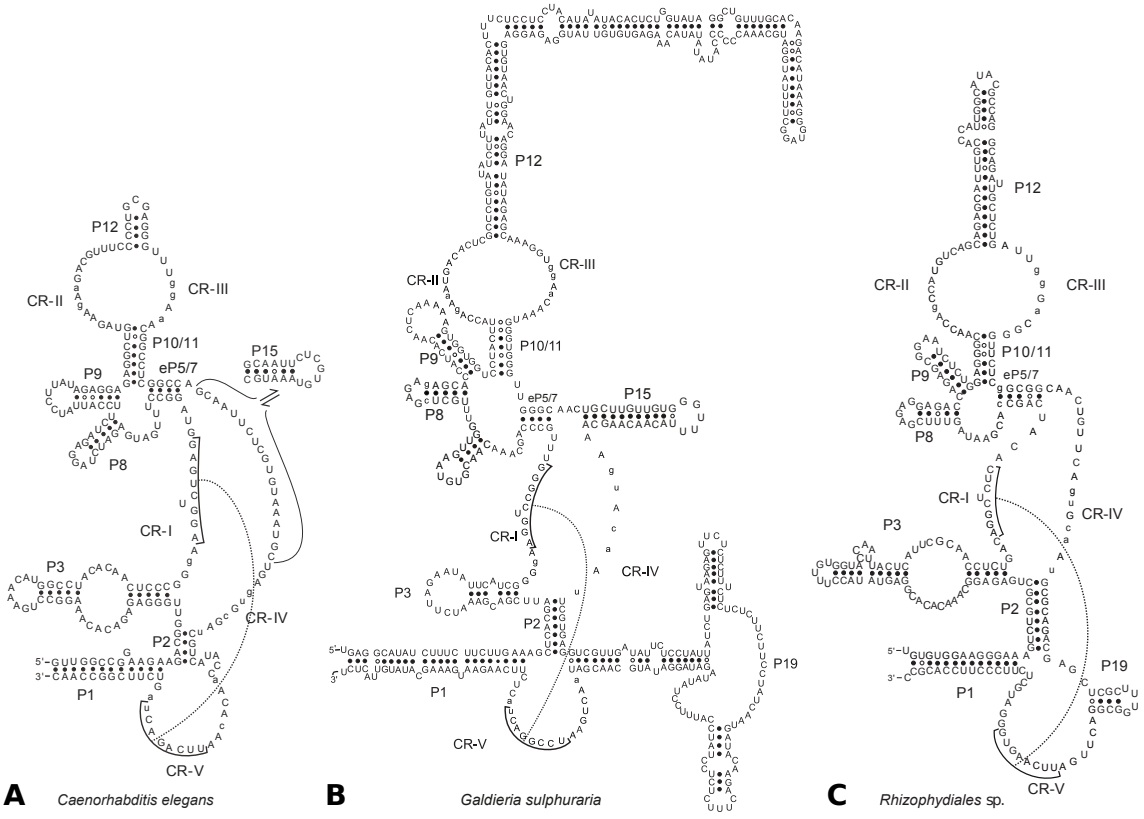


Figure 13: Predicted organellar P RNA structures. Predicted structures identified in *C. elegans*, *G. sulphuraria* and *Rhizophydiales* sp. exemplify the challenges of annotating functional P RNAs. Base-pairing regions were matched to P1–19 in the model of [49] and signature sequences from [50] checked. See Figure 5 for consensus.

The pronounced differences in predicted structures and difficulties annotating P RNAs speaks for a diversification and possible specialization within the organelle context that is probably owed to evolution of particularly bizarre tRNAs [209,210]. None of the prediction algorithms was able to identify all of the identified structures due to their diversity in different aspects. Nonetheless, the bioinformatic validation of predicted P RNA structures with multiple complementary folding and alignment tools exemplifies how a combination of different approaches is required for comprehensive results.

3.3.2 Organellar Protein-Only RNase P

Evolution has lead to increasing complexity of RNase P from bacteria, where the ribozyme retains catalytic activity without the auxiliary C5 protein *in vitro* [8], to archea where four to five protein subunits are part of the holoenzyme [47], up to eukaryotic RNase P with nine to ten accompanying proteins [211]. Some orgnanells like chloroplasts, mitochondria and plastids in plants and higher animals were shown to have lost the RNA component for tRNA processing all together.

We explored the evolution of organellar PRORP in depth by identifying homologs throughout the major eukaryote groups and comparing their phylogenetic ancestry. With the increasing amount of completely or partially sequenced genomes available, we are able to provide a more detailed look into the distribution of PRORP among the major branches of eukaryotes. The unrooted phylogenetic tree in Figure 14 gives an overview of PRORP sequences evolved in 88 representative species with multiple copies of the gene in some cases. Holozoa have evolved independently since the branching point of a common ancestor with only one variant of the protein (with the exception

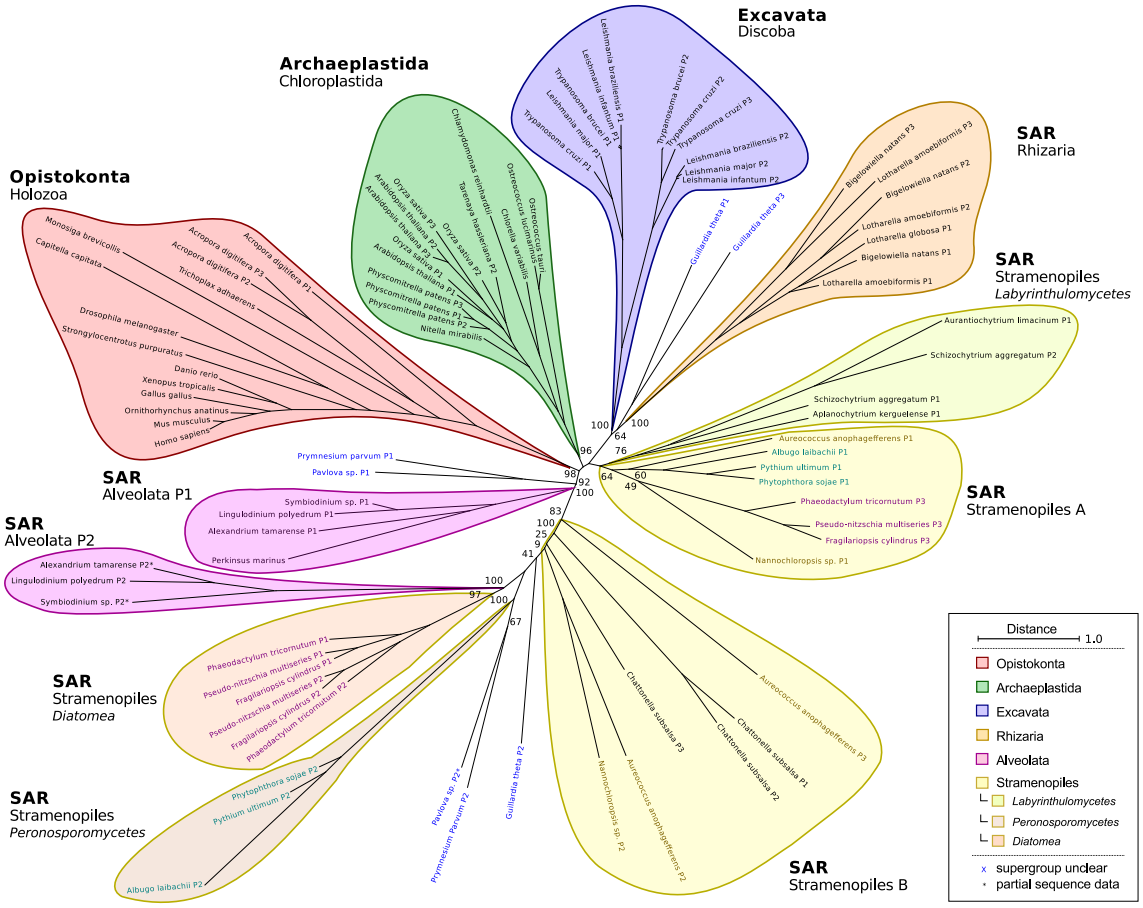


Figure 14: Phylogenetic tree of PRORP troughout Eukarya. Based on sequences of representatives from the major eukaryote groups, constructed using a maximum likelihood-based phylogenetic analysis with 100 bootstrap inferences as described in the supplementary methods. PRORP distribution suggests an ancient origin of PRORP, i.e. that it appears to have evolved in an organsim at the root of modern Eukarya, although its distribution also involved likely horizontal gene transfer (e.g. in the various Stamenophiles groups). Bootstrap support values are indicated in small numbers for the major branches. Species with unclear relation to the super groups are indicated in blue. Species with one PRORP sequence in common group Stramenopiles A and one in another distinct group are indicated in the same color.

Acropora digitifera which exhibits two duplication events within the group). Similarly, Chloroplastida developed independently with up to three copies (in *Physcomitrella patens*, *Arabidopsis thaliana* and *Oryza sativa*) only very recently in evolution. An early gene duplication event results in two distinctive proteins in Discoba with recent duplication of *Trypanosoma cruzi* P2. In contrast, evolution of PROPR among the SAR group (Stramenopiles, Alveolata and Rhizaria) is more complicated and most likely involved multiple incidents of HGT. While Rhizaria developed independently with two internal gene duplication events leading to up to 3 gene copies Stramenopiles carry up to three very distinctive homologs partially resembling a variant present in Alveolata which goes back to an early duplication event. Especially the Stramenopiles B group contains sequences only remotely related to the rest of alignment.

Since the discovery of an organellar RNase P without RNA component the concept seems to find widespread adoption among Eukarya with multiple duplication events and potential HGT that allowed for keeping pace with degenerating mitochondrial tRNA structures [55]. Even though the mere existence of PRORP homologs in these species is no sufficient evidence for the absence of an RNA component without further experimental verification this potential riddance could mark an additional step on an evolutionary path from an RNA-based [212] to a protein-centric world [213].

The powerful combination of multiple structure and homology prediction algorithms produced complementary sets of RNase P candidates to explore the diversity and evolutionary development of the underlying enzyme. However, as with any automated approach, careful re-evaluation of predictions and plausibility checking of the evolutionary implications needed to be monitored closely and if possible backed-up by multiple methods. Thus cross-checking structure predictions with other tools and statistically rigorous phylogeny were the key to the presented findings.

3.4 Identification of circRNAs in Honeybees Brains

This zoological study set out to determine the presence of circRNAs in *A. mellifera* in transcriptomic data. A further genomic characterization of circularized loci is supposed to shed light on the evolutionary conservation of this particular group of ncRNAs by comparing homologs previously identified in *Drosophila* and *B. mori*. Ultimately, the goal was to find out whether task depended differences in circRNA expression levels exist between nurse and forager bees and therefore reveal potential developmental functions of circRNAs for the first time in this unique social model.

3.4.1 circRNAs Are Detectable in Conventional RNA-Seq Data

In an initial pre-screening of publicly available sequencing data we wanted to assess whether the search for circRNAs in honeybee brains was promising before investing considerable time and money into laboratory experiments. Given the diverse spectrum and large amount of previous transcriptomic data, we were able to detect 381 BSJs with significantly many supporting JSR across most of the libraries. It must be noted that these circRNA candidates deviated largely from the circRNAs we confidently identified in our

own specifically enriched data. The most abundant circRNA in this non-specific dataset is *ame_circ_0002566* with 8,479 JSRs across only a small subset of the libraries, while only 14 reads supporting the junction were picked up in our sample. The second most abundant BSJ present in almost all libraries is *ame_circ_0000163* which is also fairly abundant among our enriched libraries with 103 JSRs. Some libraries, especially those from Project `PRJNA200755`, `PRJNA257666` and `PRJNA227348` did yield only few or no JSRs, possibly due to their preparation and selection protocols, which were generally not suitable for this type investigation. Due to the large number of junctions reads in public data that is not present in the specifically enriched libraries of our study we suspect a considerable amount of false positives particular to individual datasets and over-sensibility of the `segemehl` mapping software. To avoid over-estimation in our own experiments we thus applied extremely conservative constrains to list of reported circRNAs and employed an identification pipeline with two independent mapping and calling implementations in parallel.

Despite some of the potentially conflicting results of this initial screening we were able to show that evidence for circRNAs can be found without specific experiments given public data analogous to [68]. Even though we were able to confirm some of the detected BSJs later on, one has to be mindful of false positives or easily draw quantitative conclusions from external data.

3.4.2 RNase R Enriches Circular Transcripts

We prepared RNA-Seq libraries from total RNA extracts of honeybee worker brains enriched for circular RNAs and compared them to a non-enriched library. Each BSJ supported by multiple JSRs was considered as representative of a distinct circular RNA. We were able to detect a total of 3,384 individual BSJs supported by at least three JSRs from the four libraries combining two different methods, see Figure 15A. Based on these we provide two sets of circRNAs identified by applying different stringency thresholds. The low-stringency sets contains 1,263 circRNAs found by both independent algorithmic

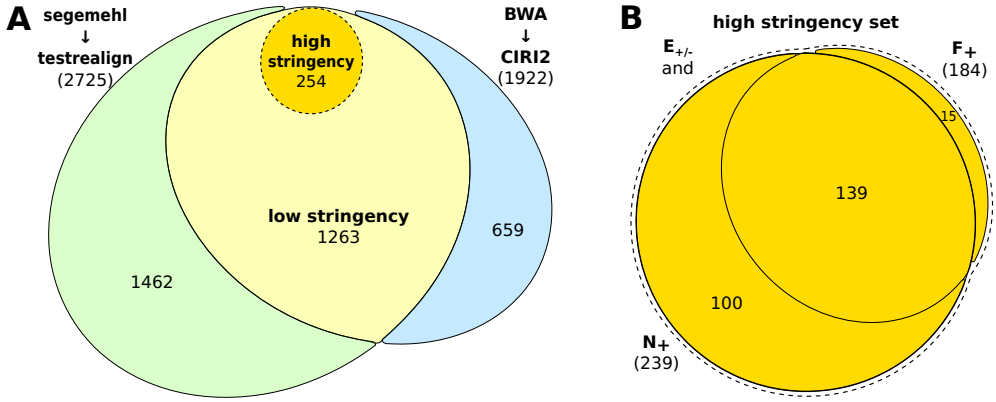


Figure 15: Identified circRNAs by RNA-Seq. (A) Two independent algorithms were used to predict circRNAs: `segemehl` in conjunction with `testrealign` (green) and `BWA` with `CIRI2`. The overlap was considered as low-stringency set (yellow). The high stringency set additionally requires an enrichment through RNase R treatment and compelling read coverage from at least two independent sequencing libraries. (B) Strong evidence is found in three independent sequencing libraries for 139 candidates. These candidates were also identified as enriched in E_+ vs E_-.

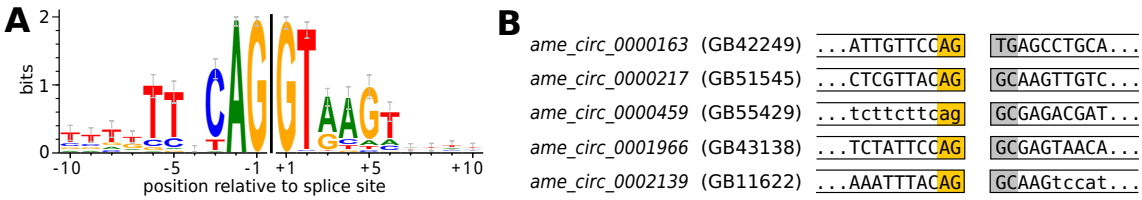


Figure 16: Splice site motif in flanking introns of circRNAs. (A) Flanking introns of circRNA junctions exhibit a fairly common `ttxcAG/GTaagt` motif. (B) Among our high confidence circRNAs, only five showed non-canonical `AG/GT` junctions which, however, correlated with annotated exon boundaries. Acceptor sequences (yellow) are canonical while the donor sequence (gray) deviates in these cases. Small letters indicate regions marked for low-complexity within the overall genomic context.

methods (overlap). Only these BSJs were considered viable circRNA candidates because previous studies showed inconsistent results between different algorithms [83,164]. Specifically, `segemehl` is known to produce very sensitive mapping results, potentially introducing false positives when solely relied upon [164]. The high stringency set is a subset containing 254 circRNAs with a higher amount of supporting reads along with a significant five-fold enrichment of the JSRs through RNase R treatment. The majority ($>$ 77 %) of the circular transcripts were even enriched by more than ten-fold. We remark, that these numbers refer to circRNAs that are expressed in the brain of nurse and forager bees. In contrast, 2,513 circRNAs reported for *D. melanogaster* [68] and 3,916 for *B. mori* [87] are based on samples of different developmental stages, tissues and even cultured cells and do not ensure RNase R enrichment.

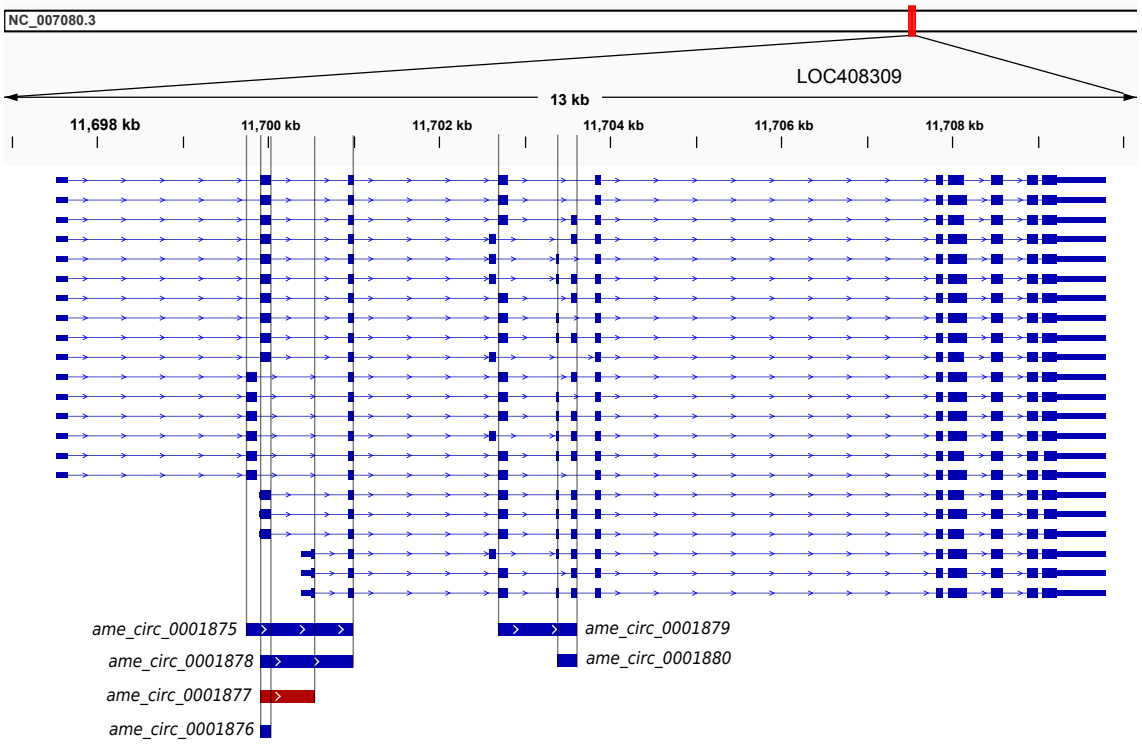


Figure 17: Ambiguous transcript assignment. There is currently no annotated isoform of `LOC408309` which includes both exons of circRNA *ame_circ_0001877*. Both exons are part of a different 5'-UTR and end with a coding region. The circRNA is well supported with 121 JSRs and likely a result of an excision from one of the longer isoforms of the transcript that contain neither of the two exons (e.g. `XM\_006566436.2`).

Almost all of the BSJs are flanked by a canonical GT/AG splice signal, summarize in the sequence motif in Figure 16A. Only five circRNAs did not show such a canonical splice site (see Figure 16B). In one case proper transcript attribution was not possible because the BSJ spans two exons that (presumably) do not occur in the same isoform (illustrated in Figure 17). The coding exon of gene CG45167 (homolog of B52 in *Drosophila*) and an immediate downstream exon which starts with a different 5'-UTR are not present in any currently annotated transcript variant. Such an example could indicate that back-splicing is indeed due to erroneous alternative splicing events.

The amount of canonically spliced transcripts (linear) is at least the same as the amount of back-spliced transcripts (circular) for the majority of circRNAs identified here. For this reason it is unlikely that the circRNAs presented here arose from a mapping artifact, e.g. due to misalignment of reads or repeating gene copies because exons were splice either way.

A complete list of all detected circRNAs including read levels and putative homologs in *Drosophila* and *Bombyx* is part of the manuscript in Section 5.4. An excerpt of the most prominent entities is shown in Table 7.

Table 7: Excerpt of identified circRNAs in the brain of honeybee nurse and forager bees. All circRNAs were significantly enriched in E_+ over the non-enriched set E_-. The respective chromosome is indicated in the **Chr.** column. The summarized number of **JSRs** is given along with the averaged normalized expression (**Expr.**) and fold enrichment (**Enriched**). The **Homology** column indicates whether a *Drosophila* or *Bombyx* homolog was found in ⋆ [73], ◦ [68] or † [87]. The full list can be found as part of the submitted manuscript.

circRNA ID	Host gene	BeeBase	Chr.	JSRs	Enriched	Expr.	Homology
ame_circ_0001970	*LOC413427*	GB43145	LG11	432	6.7	0.329	⋆◦
ame_circ_0000721	*LOC724885*	GB53835	LG3	139	10.2	0.344	
ame_circ_0002142	*LOC410393*	GB52063	LG12	124	5.0	0.628	†
ame_circ_0000163	*LOC408576*	GB42249	LG1	103	7.2	0.328	⋆◦
ame_circ_0000232	*mup2*	GB49259	LG1	97	11.0	0.119	
ame_circ_0001780	*rad*	GB49511	LG10	91	5.0	0.062	⋆◦
ame_circ_0001286	*LOC411534*	GB44365	LG7	63	93.0	0.731	⋆◦
ame_circ_0002579	*LOC409655*	GB47584	LG16	42	17.5	0.226	
ame_circ_0001822	*rsmep2*	GB54272	LG10	34	5.0	0.022	⋆◦
ame_circ_0002577	*LOC409655*	GB47584	LG16	17	11.3	0.072	
ame_circ_0001852	*coRest*	GB52614	LG10	10	5.0	0.013	⋆◦
ame_circ_0001099	*LOC411114*	GB44582	LG5	306	18.5	0.328	†
ame_circ_0000414	*LOC725294*	GB55364	LG2	216	7.6	0.312	⋆◦
ame_circ_0000397	*LOC408688*	GB49767	LG2	185	5.0	0.377	⋆◦†
ame_circ_0001712	*LOC408996*	GB42579	LG9	169	6.3	0.198	⋆◦†
ame_circ_0002576	*LOC409655*	GB47584	LG16	168	14.9	0.644	
ame_circ_0001638	*LOC411347*	GB17597	LG9	159	9.4	0.400	
ame_circ_0001593	*LOC408991*	GB53310	LG9	148	7.9	0.105	
ame_circ_0001479	*LOC408957*	GB40504	LG8	147	18.4	0.339	◦
ame_circ_0000524	*LOC408718*	GB43446	LG2	130	36.2	0.313	
ame_circ_0001120	*sGC-alpha1*	GB52929	LG6	129	10.4	0.276	†
ame_circ_0000054	*LOC726544*	GB42188	LG1	124	7.5	0.480	
ame_circ_0001877	*LOC408309*	GB45167	LG11	121	9.4	0.085	
ame_circ_0000669	*LOC410044*	GB55791	LG3	118	13.0	0.370	
ame_circ_0000073	*LOC410717*	GB55293	LG1	111	11.0	0.290	⋆◦
ame_circ_0001340	*LOC411229*	GB42567	LG7	109	6.9	0.570	◦

Complementary to the BSJs detected by RNA-Seq, circularity of predicted loci was verified with independent PCR experiments. Figure 18 shows PCR products of circRNAs amplified with convergent and divergent primers listed in Table 4 and *ef1α* as a stably expressed control gene. Given, that these circRNA candidates can be verified so cearly with two independent methods, creates confidence in the RNA-Seq results and reliable enrichment through RNase R in general. Basing further analysis on the set of high-stringency circRNAs becomes therefore justifiable.

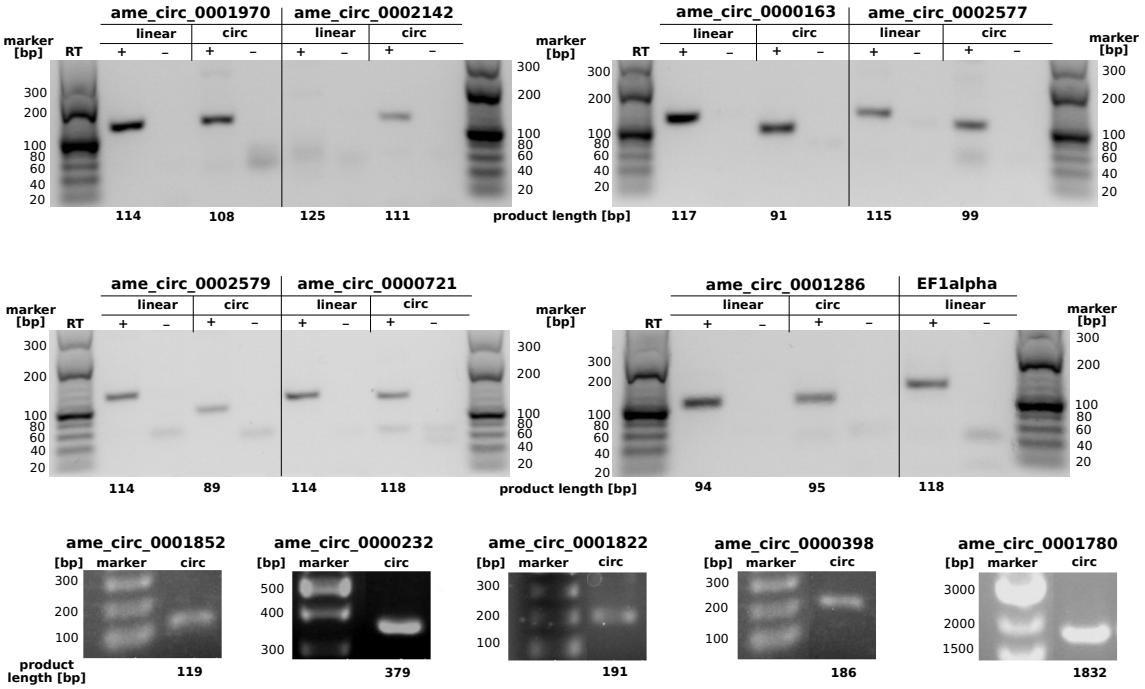


Figure 18: PCR amplification of circRNA with divergent primers. A selection of circRNA candidtes identified in RNA-Seq experiments were validated with convergent (linear) and divergent (circ) primer pairs spanning the back-splicing junction. Samples of total RNA with (+) and without (-) reverse transcription into cDNA were compared in order to destinguish genomic DNA amplification. Anticipated PCR product lengths are indicated below the gels. Note that the PCR products run slightly higher than indicated by the leader because we prestained with GelRed. For the five bottom BSJs RT(-) control and linear amplicons are not shown.

3.4.3 circAm*rad* Shows Task Dependent Expression

The circRNAs *ame_circ_0001780* and *ame_circ_0001822* showed a notable differential expression pattern in RNA-Seq results of nurse bees and foragers. For simplicity they will be termed according to their host genes in the further course of the study: circAm*rad* and circAm*rsmep2*, respectively. As the experimental setup is not suitable for any reliable quantitative assertions, we decided to perform a targeted qPCR for these candidates at different developmental stages. In addition, we compared the expression patterns in bees with age-related task allocation to those undergoing a task allocation due to colony needs (age-unrelated, SCC), see Figure 19.

For circAm*rsmep2* we found that expression in the brain is higher in foragers than in nurse bees (Figure 19A). This difference, however, does not seem to be directly task-related. In a SCC where nurse and forager bees have exactly the same age, no expression differences are observed (Figure 19A). Our interpretation is that this expression difference most

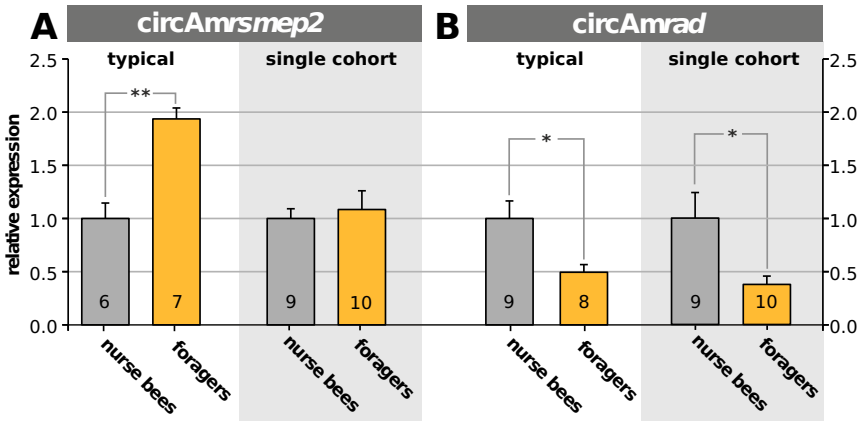


Figure 19: Quantitative expression analysis. TaqMen expression analysis of (A) circAm*rsmep2* and (B) circAm*rad* in brains of nurse bees and foragers from colonies with a typical age structure and SCCs consisting of bees of the same age. Expression is given relative to nurse bees. Bars show relative mean expression with standard error. The number of replicates are indicated in the bars. Significant differences are indicated ($\star p < 0.05$, $\star\star p < 0.01$, Two tailed unpaired Students t-test). In contrast to circAm*rsmep2* the expression of circAm*rad* seems correlated with the allocated task rather than the bees' age.

likely depends on the bees' age but not on its task. Supposedly, circAm*rsmep2* accumulates over time in the brain of worker bees, as shown for certain circRNAs in the nervous system from mammals to flies [68,214]. On the other hand, a significant increase of the linear product in foragers was reported previously (`XM_393489.3`/Am*rsmep2*, log_2ratio ~2.8) [215]. The observed increase of the circular product circAm*rsmep2* might thus be a consequence of generally increased expression of the host gene, which codes for a RIM-family (Rab3a-interacting molecule) protein. Studies in species of the tetrapoda clade (human, mouse, chicken and so on) show that this family plays an important role in neuronal plasticity, especially in neurotransmitter release and in organizing active zones in plasma membranes [216,217].

In contrast, circAm*rad* is higher expressed in brains of nurse bees than in brains of foragers (Figure~19B, typical). Strikingly, this is inversely correlated with the expression of the linear product which is strongly increased in foragers (`XM_393494.2`/Am*rad*, log_2ratio ~6.1) [215] and holds true independent of the age-related task transition. The expression levels in the SCC experiment (Figure 19B) are similar to that of typical colonies where tasks are allocated based on a bee's age. This data suggests a correlation of acquired task and circAm*rad* levels. Either the task of the bee is influencing circAm*rad* expression or *vice versa*. Its host gene is orthologous to the *radish* gene in *D. melanogaster*, which is known to play a crucial role in the amnesia-resistant memory (ARM). Unlike the long term memory ARM does not require protein *de novo* synthesis [218] and thus represents a low-costs memory form [219,220]. *Rad* also possesses circRNAs in fly (Table 7), but whether this circRNA is involved in ARM or whether ARM is also present in honeybees, has not yet been investigated.

3.4.4 Circularization of Exons Is Evolutionarily Conserved

Conservation of loci is well-known between mouse and human circRNAs and it therefore stands to reason that circRNAs in honeybees can also be found in *Drosophila* and *Bombyx*. Honeybee circRNAs were compared to those found in fruit fly [68,73] and silkworm [87]

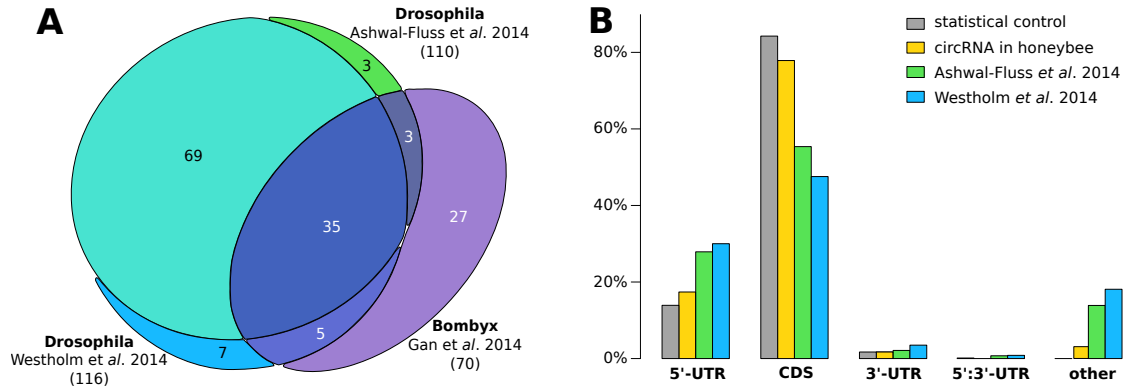


Figure 20: Homology of circRNAs. (A) 122 host genes are orthologous to host genes of circRNAs identified in either *Drosophila* or *Bombyx* in previous studies. (B) BSJs in honeybees and fruit fly are clustered into the following categories: part of the 5'-UTR, from the coding sequence exclusively, part of the 3'-UTR, spanning from 5'- to 3'-UTR or other (in the middle of exons, introns, part of non-coding genes or in between genes).

based on homology of their parental genes (Figure 20A). Out of 254 honeybee circRNAs only 70 host gene homologs were found in silkworm (30 %). In contrast, 203 homologs were identified for fruit fly (80 %) which can be explained by the closer phylogenetic relationship to honeybee [221]. Consistent with our results, circularized exons in fruit fly were found in 144–151 of these homologs (with respect to [73] and [68], overlap of 122). This finding is in line with a similar comparison of circRNAs in human and mouse [86]. There, two thirds of all host genes harboring back-splicing junctions could be correlated by homologies between the two species. Even though circRNAs are known for only three insects so far, the number of homologous host genes among them suggests that circRNAs are commonly found in insects. Features identified for circRNAs in one of these organisms are likely to be valid for other insects.

Being able to compare circRNA in different insects also permitted further characterization of which exons in a transcript are statistically prone to circularization. The majority of BSJs in honeybee correspond exactly to exon boundaries of protein coding regions (78 %), see Figure 20B. Nearly all remaining cases are derived from 5'-UTR containing segments (17 %). This is only slightly different from the set of (presumably) linearly spliced exons in the control but shows a trend towards 5'-UTRs. For both *D. melanogaster* datasets [68,73] the overall proportion is similar but with a much stronger bias towards 5'-UTRs (~30 %) and non-canonical splice events, e.g. occurring in the middle of introns or exons in between genes (other: ~20 %). The latter category was rarely found for honeybee circRNAs ($\leq$ 2 %). We note that this difference might be a result of different annotation

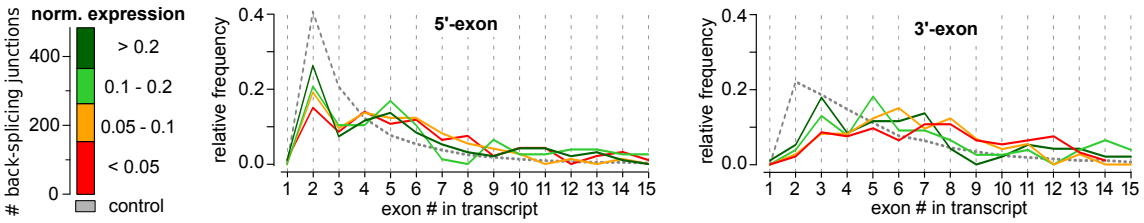


Figure 21: Exon position of circRNAs in the host gene. First exon at the 5'-end the circRNA (left) and last exon at the 3'-end (right) within the host gene, by number of exons. CircRNAs are stratified by normalized circRNA expression into four groups (green to red with decreasing ratio). As reference, randomly selected exons in the control (gray) exhibit a higher number of second and third exons in a transcript just by chance.

qualities for honeybee (data from 2018) and fruit fly (data from before 2014) and should thus not be over-interpreted.

For fruit fly it was reported that circRNAs mostly spring from the second exon of a transcript [68]. This is also true for honeybee circRNAs. Figure 21, however, shows that this number is implied by the outstanding abundance of transcripts with only two exons. This is also visible in the randomized control distribution. Compared to this set, the observed starts at exon two are actually less than what would be expected. We identified two factors that correlate with back-splicing: 1) The exon position. 2) The number of exons. The further downstream an exon is located in a transcript and the more exons (and thereby splice-junctions) it exhibits, the more likely circRNAs arise from the transcript.

3.4.5 Correlation with Memory-Associated Loci

A GO term analysis (gene ontology term enrichment) was performed using all 203 circRNA host gene homologs correlated to fruit fly from which we extrapolated the functional annotation. An excerpt of enriched terms is shown in Table 8. High level processes involved in synaptic development and regulation were significantly enriched. Given that the source samples were obtained from brain tissue, this is an expected result but it also resembles the finding that neurologically associated genes are a main source of circRNAs as found for *Drosophila* [68].

The most enriched high level terms below a p-value of 10^{-4} were 'anesthesia-resistant memory' (27x), 'medium-term memory' (23x), 'regulation of neuromuscular synaptic transmission' (21x) and 'deactivation of rhodopsin mediated signaling' (21x). The former is especially remarkable. One representative of this group is the *radish* gene from which the circRNA circAm*rad* (*ame_circ_0002363*) arises. We found its abundance levels in correlation with the acquired task of a bee (see the manuscript in Section 5.4).

Table 8: GO term enrichment of biological process among circRNAs. Functional annotation is based on homology to *Drosophila* genes and statistically compared to all reference genes in the PANTHER database. Only the top level terms with five fold enrichment in circRNA associated genes (circ) over the reference set (ref) above expected (exp) and sufficient significance ($p < 10^{-4}$) after multiple testing correction are presented.

biological process	GO term	circ / ref (exp)	Fold-enrichment	p-value
anesthesia-resistant memory	0007615	4 / 13 (0.15)	26.79	3.4×10^{-5}
medium-term memory	0072375	4 / 15 (0.17)	23.22	5.4×10^{-5}
reg. of neuromuscular synaptic transmission	1900073	5 / 21 (0.24)	20.73	9.6×10^{-6}
deactivation of rhodopsin mediated signaling	0016059	4 / 17 (0.2)	20.49	8.2×10^{-5}
potassium ion transport	0006813	7 / 43 (0.49)	14.17	1.4×10^{-6}
synaptic growth at neuromuscular junction	0051124	6 / 41 (0.47)	12.74	1.4×10^{-5}
neuromuscular synaptic transmission	0007274	7 / 57 (0.65)	10.69	7.7×10^{-6}
myotube differentiation	0014902	6 / 49 (0.56)	10.66	3.6×10^{-5}
synapse assembly	0007416	7 / 70 (0.8)	8.71	2.6×10^{-5}
regulation of membrane potential	0042391	7 / 75 (0.86)	8.13	4.0×10^{-5}
compound eye photoreceptor development	0042051	7 / 86 (0.99)	7.09	8.9×10^{-5}
response to light stimulus	0009416	11 / 151 (1.73)	6.34	2.3×10^{-6}
synapse organization	0050808	12 / 174 (2)	6	1.4×10^{-6}
regulation of synapse organization	0050807	9 / 142 (1.63)	5.52	5.6×10^{-5}
taxis	0042330	21 / 332 (3.81)	5.51	5.2×10^{-10}
protein phosphorylation	0006468	17 / 279 (3.2)	5.31	4.4×10^{-8}
locomotory behavior	0007626	11 / 186 (2.14)	5.15	1.5×10^{-5}
response to drug	0042493	12 / 205 (2.35)	5.1	6.8×10^{-6}

Consistent with this is also the enrichment of rhodopsin signaling and memory-related genes. Nurse bees take care of the brood inside the hive, where it is dark and the requirements to memory are different from those of foragers [222]. After task transition to forager bees, they start to collect food from outside the hive, mostly at daylight, and need to find their way back to the hive afterwards. A need for adaptation of rhodopsin signaling and a change in memory requirements is obvious. In fact, 'positive phototaxis' showed the highest GO term enrichment (44x). The p-value however was below the applied threshold ($1.87_{\times 10^{-3}}$) because the term only has four representatives in the reference set. A overview of most enriched GO terms can be found in Table 8.

Despite the neural origin of total RNA used for the identification of the circRNAs, the functional over representation of memory-associated host genes is remarkable. Possibly, a correlation with increased host gene expression can be the cause of this enrichment. However, many of the characterized circRNAs exhibited converse relative expression levels between the two castes and previous studies reported that expression levels are independent from host gene levels [72]. Finding the largest abundances of circRNAs in neural tissue might thus be a product of their neuro-develpmental function.

3.4.6 Increased miRNA Targets in Conserved circRNAs

Some of the previous circRNA studies in humans and *Drosophila* reported an increased amount of miRNA binding sites in the circularized exon sequences [61,68,85] while others could not come to the same finding [71,76,86]. Thus, potential miRNA target-sites were annotated for all 254 high-stringency circRNAs identified in honeybee and analyzed for statistical enrichment. The results can be divided based on their degree of phylogenetic conservation as seen in Figure 22. $3,058$ target sites were only conserved in *Apis* species. We argue that *Apis* species are too closely related to qualify as reliable predictor for miRNA target sites. The sequence conservation in this set appears rather high in general. This is also reflected by a similar distribution of potential miRNA target sites compared to the control without any constraints on conservation, see Figure 22A.

A set of $1,076$ sites is conserved in *Apis* and eusocial insects which are sufficiently distant to *A.~mellifera* to reasonably infer conservation. With about 10.4 target sites per $1,000$ nt circRNAs have a 1.7x increase in miRNA target sites compared to the median of linear splice product control. Thus, in line with previous findings for *Drosophila* [68], we report a general enrichment of conserved miRNA target sites in circRNAs over random linear counterparts. The most enriched miRNA target sites correspond to *ame-miR-3748/ame-miR-3753* (~10x enriched, same seed region) and ame-miR-3791 (~9.2x enriched), see Figure 22B. RNA expression studies show that the abundance levels of some miRNAs correlate with task or age of honeybees [223–226]. We did, however, not find a significant overlap of miRNAs corresponding to enriched target sites and miRNAs reported as differentially expressed between nurses and foragers in these studies. A complete list of potential target sites and their degree of conservation is part of the manuscript in Section 5.4.

The differences in statistical over representation at various conservation levels show that an interpretation of these results is highly dependent on the method of miRNA target

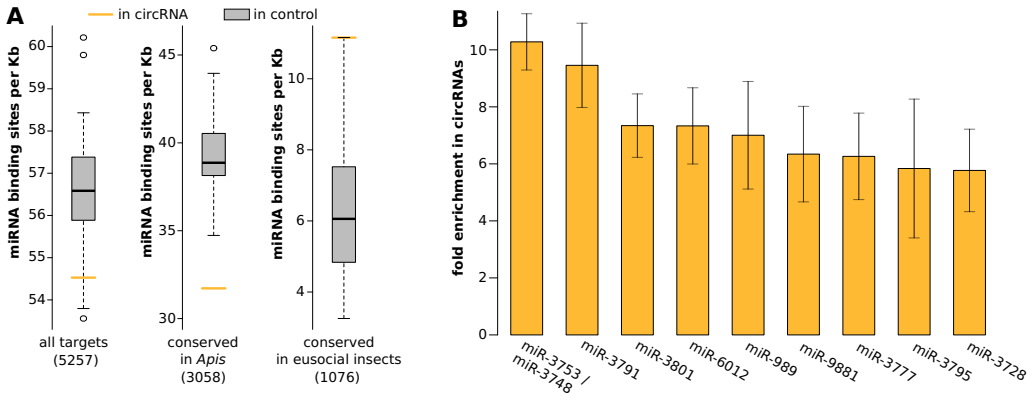


Figure 22: Putative miRNA target sites in circRNA exons. (A) Putative target sites normalized by exon length in differentially conserved sets. 'All targets' represents no conservation at all. *Apis* represents conservation only in closely related species. Eusocial insects are sufficiently distant to consider conservation relevant. The number of miRNA binding sites conserved in this set is significantly higher ($p < 0.001$, Student's t-test) than in the control. The absolute number of potential binding sites in the respective sets is given in parentheses below. (B) Illustration of miRNAs with potential target sites in circRNAs conserved even in eusocial insects. Only miRNAs with at least ten target sites and an at least 5-fold enrichment over control are shown. Error bars indicate standard deviation in fold enrichment between different controls (N= 42).

prediction. Discrepancies between previous studies can therefore probably explained because conservation was regarded or not. Available tools, often do not provide this extra step for target prediction or are only suited for human data [206]. Based on the statistical data presented here, a general over representation is visible but none of the characterized circRNAs seems to be targeted by any specific miRNA, that would make it a potential miRNA sponge.

3.4.7 No Significant Complementarity in Flanking Introns

In honeybee, introns flanking circularized exons are significantly longer than those from linearly spliced exons, see Figure 23. They can span several thousand bases. This result is in line with findings from fruit fly and human [68,227]. There in addition, flanking introns showed increased levels of reverse complementarity compared to linearly spliced exons. Reverse complementary regions are thought to enhance the likelihood for base-pairing between the introns. This interaction likely guides back-splicing process [228–230]. Following up on this assumption, introns were reciprocally scanned for reverse complementary matches at sequence-level using BLAST [138], see Figure 24A. While the result shows that introns flanking circularized exons are composed of regions with

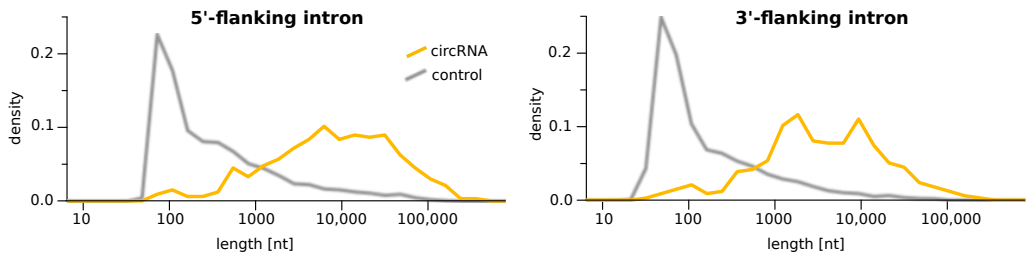


Figure 23: Introns flanking circRNAs. Flanking introns upstream (left) and downstream (right) of circRNAs (yellow) are significantly longer (t-test with $p < 0.001$ for both) than those of a random control (gray).

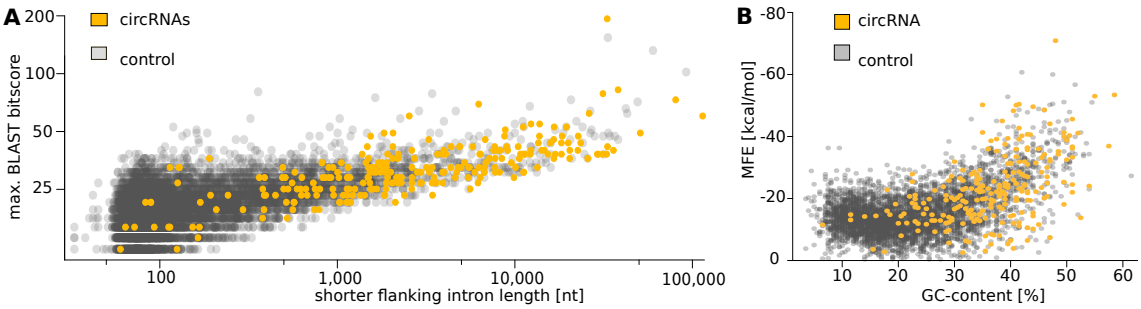


Figure 24: Introns flanking circRNAs. (A) BLAST bitscores of the best reverse complementary match of the shorter circRNA flanking intron to the other (yellow) compared to control intron pairs (gray). (B) Relation of GC-content and MFEs. Both measures appear to be linked (correlation $\rho = 0.5$, Spearman's rank correlation).

better complementary (represented by higher bitscores) in general, it is also obvious that complementarity is linked to the length of introns. Higher scores of complementarity matches are likely a result of the fact that introns flanking circularized exons are much longer than those from the control set. The most relevant regions for circularization are probably the end of the 5'-flanking and the start of the 3'-flanking intron.

An RNA secondary structure prediction using RNAfold [184] was used to investigate potential intron-intron interactions more specifically, see Figure 24B. The difference is more obvious using this method. Co-folded complexes of the control introns exhibit much higher MFE scores, indicating less base-pairing interaction. However, the MFE scores partly cover similar ranges, which does not allow for a clear distinction between circularized exons and linear splicing products. The graph also shows that the increase in folding potential (represented by lower MFE scores) is linked to GC-content of the respective introns. Also the fact that the complementarity match as well as the co-folding analysis yielded similar results for all combinations of starts and ends of the flanking introns (e.g. pairing the end of the upstream intron with the end of the downstream intron) puts a direct effect of base-pairing in doubt. The GC-content in turn well discriminates circRNA introns from control introns, see Figure 25B.

3.4.8 Increased DNA Methylation in Flanking Regions

The intronic features raise the question, why the GC-content of circRNA flanking introns is elevated in such significant amounts (median shifted from 20 % to 36 %, $p < 0.001$). One reasonable explanation is an increase of potential DNA-methylation at these introns due to CpG islands. While the exact mechanism is unknown so far, DNA-methylation is known to induce alternative splicing in honeybees [103,231]. Methylation patterns also vary depending on the age and allocated task of an individual bee [101,232,233]. It was even shown that reverted nurse bees regain their original methylation patterns independent of their age [92]. Figure 25C illustrates that the CpG dinucleotide frequency is also significantly increased for circRNA flanking introns and nearly absent in the control group (~1 %). As CpG sites are preferentially methylated [101,232], this indicates a significant increase of potential DNA-methylation sites. Moreover, cytosine methylation and hydroxymethylation at non-CG sites (CA, CT, CC) is reported to be enriched in introns of the honeybee [234]. In line with this, Figure 25D shows that also the cytosine

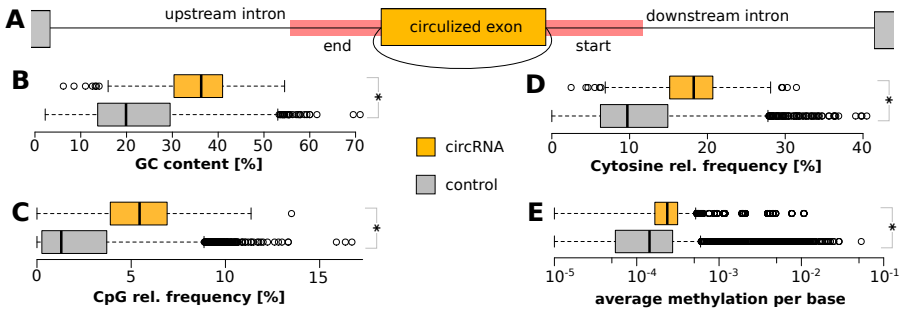


Figure 25: Putative interactions and sequence properties of flanking introns.(A) Scheme indicating the relevant regions (red) of introns flanking circularized exons (yellow). (B) Screen of GC-content. A significant increase is found compared to the control. (C) Screen of CpG dinucleotides. A significant increase is found compared to the control which rarely exhibits CpG. (D) Screen of cytosine-content. A significant increase is found compared to the control. It explains the observed effects for GC-content and CpG frequency. (E) Screen of average base methylation. The number of methylated bases in introns flanking circRNAs is slightly but significantly increased compared to linear control introns. $\star p$-value < 0.001

mononucleotide frequency is significantly increased for circRNA flanking introns. While the genome comprises ~16 % cytosines, circRNA introns exhibit a median of ~18 % cytosines. Strikingly, the median cytosine-content of linearized exons is as low as 10 %. This can be translated into reduced methylation and hydroxymethylation potential and thereby fewer alternative splicing events for introns flanking canonically spliced RNAs compared to those that frequently result in circRNAs.

We evaluated publicly available whole genome bisulfite sequencing data of worker bees from a previous study to comprehensively determine methylation levels [92]. Figure 25E shows that the length-normalized accumulative DNA-methylation of introns flanking circular RNAs actually tends to be increased compared to those flanking random exons. Notably, the effect was not visible using only the closest 50 or 100 nucleotides of a flanking intron but became visible using a 200 nt window or full-length introns. This is probably due to the limited windows size which is likely too small for statistical assessment.

While relevant social roles in the used methylation study [92] are the same, we note that collection times and extraction methods differ from experiments done in this study. Ideally, the libraries used for circRNA detection and DNA-methylation analysis should be derived from the same biological sample. Without further experimental investigation a strong conclusion cannot be drawn yet. We argue however, that the data presented here provides first indications for a link of circularization and DNA-methylation in honeybees. It is possible that the age-depended increase of circRNA abundance is not (only) due to potentially lower decay rates of circRNAs compared to linear products but also a result of increasing DNA-methylation that leads to alternative splicing accompanied by increase of circRNA formation.

4 Conclusion & Outlook

Bioinformatics has become an indispensable discipline in molecular biology. The major parts of this thesis have in common that biological data is infeasible to interpret by hand and thus computer aided methods have to be applied to answer the original research question. Transcriptomic data is easily and cheaply generated nowadays or even publicly available from other studies. However, the outcome of their analysis strongly depends on the bioinformatic approach applied to their interpretation. These novel approaches require state-of-the-art algorithms to cope with the magnitude of modern high-throughput experiments in order to process them in a feasible time frame without high performance computers. Throughout the projects presented in this thesis, some of the most recent software packages were used, because only highly optimized algorithms can handle the massive data of the often specialized experimental setups, such as split-read mapping for circRNAs. Because experimental protocols evolve constantly in the ever progressing field of molecular biology, developing custom tailored bioinformatic solutions on top of existing algorithms in tight cooperation with experimental researchers is a requirement to assure case-specific interpretation of the results.

This work presents several examples of finding supporting evidence for ncRNAs in publicly available data of the same and related species. In order to rule out, that the unusually long 5'-UTR of the DCW cluster and its embedded sRNA are not a singular artifact in *R. sphaeroides* alone, it was imperative to compare transcriptomic data in related taxons. Finding a similar feature in closely related species corroborated our characterization of an transcriptional regulator in this evolutionary sub-branch of rod-shaped bacteria. Similarly, the rapidly growing amount of fully sequenced genomes facilitates phylogenetic comparisons of unprecedented detail as seen in the evolutionary exploration of organellar RNase P. Even though we relied on our own experimental data in the identification and quantification of circRNA in honeybees, a previous screening for BSJs in non-enriched public data bootstrapped our investigation and the already available bisulfite data enabled the confirmation of increased methylation surrounding circularized exons. This aspect of the thesis also highlights the strength of open and shared research data inside a fast evolving research field and the possibility to reuse

expensively generated resources in order to circumvent redundant experiments, save time and confirm results based on independent data.

Besides the combination of different datasets, making use of multiple algorithmic approaches in conjunction provides either a more comprehensive understanding of the results or might show off biases one would have encountered when only relying on one method. In the attempt to determine riboswitch functionality of UpsM, exploring the landscape of possible suboptimal structures with barrier trees did not reveal any switching mechanism, while manually constraining the 3'-end lead to convincing refolding of the ncRNA. Similarly the annotation of RNase P RNAs relied on the complementary predictions of multiple structure alignment implementations to offer maximal sensitivity. The combination of two circRNA detection methods significantly increased the confidence in the reported set of candidates as the different biases of both approaches canceled each other out. In a fast-evolving scientific sphere, it is therefore beneficial to consult multiple cutting edge approaches in parallel as long as a gold standard is not established yet.

Each of the studies set out with a biological research question that could best be answered with sequencing approaches where applied bioinformatics are vital for the interpretation of large amounts of produced data. Only in-depth bioinformatic analysis of differential transcriptomic data revealed the exact TSS of UpsM. Further amounts of large datasets then enabled the complete characterization of this conserved ncRNA in a structural alignment of homologs to disclose a possible structural rearrangement for transcription regulation. In the case of 6S RNA the influence of different RNases on its degradation could only be pinpointed through differential RNA-Seq. The visualization of processed transcripts in paired-end data is thus a unique approach as it makes the effects of each RNase easily comparable. Lastly, the genome-wide analysis of circRNA flanking intron sequences and statistical observation of a potential link between methylation status and back-splicing activity would not have been possible without the implementation of custom tailored scripts and modeling of a statistical control. Following up on such sideline observations increases the serendipity that often leads to unexpected discoveries. Additionally underpinning potential miRNA target seeds with conservation of the gene region in related species revealed that the quality of predictions can greatly alter conclusions that can be drawn from statistical inferences. In most cases, readily available solutions to common sequencing-related problems in molecular biology exist today. However, in order to interpret their results for a given research questions, it is advisable to spend the resources to develop custom, case-specific solutions on top of existing ones.

All of these solutions were developed with re-usability in mind to enable other researchers to apply the same pipeline to their data or extend these tools to their needs. The used pipelines, tools and custom scripts are available with source code and documentation in publicly available repositories. Releasing the data processing pipeline along with the data also enables other researchers to reproduce and validate the presented results.

5 Research Articles

The following scientific articles are part of this doctoral thesis and contain some of the results presented in the Results Section. My contribution to these projects and the publication status at the time point of submission of this thesis are stated before each manuscript/article.

Weber L, **Thölken C**, Volk M, Remes B, Lechner M, Klug G (2016) *The Conserved Dcw Gene Cluster of R. sphaeroides Is Preceded by an Uncommonly Extended 5' Leader Featuring the sRNA UpsM*. PLoS One; doi:10.1371/journal.pone.0165694

Damm K, Wiegard JC, Bach S, Lechner M, **Thölken C**, Hain T, Ngo S, Putzer H, Hartmann RK (2018) *Processing and decay of 6S-1 and 6S-2 RNAs in Bacillus subtilis*. (manuscript in preparation)

Lechner M, Rossmanith W, Hartmann RK, **Thölken C**, Gutmann B, Giegé P, Gobert A (2015) *Distribution of ribonucleoprotein and protein-only RNase P in eukarya*. Molecular biology and evolution; doi:10.1093/molbev/msv187

Thölken C, Thamm M, Erbacher C, Lechner M (2018) *Sequence and structural properties of circular RNAs in the brain of honeybees (Apis mellifera)*. BMC Genomics; (under review)

5.1 The Conserved Dcw Gene Cluster of *R. sphaeroides* Is Preceded by an Uncommonly Extended 5' Leader Featuring the sRNA UpsM

Authors: Lennart Weber, **Clemens Thölken**, Marcel Volk, Bernhard Remes, Marcus Lechner and Gabriele Klug

Journal: PLoS One

DOI: 10.1371/journal.pone.0165694

Contributions: Transcriptional analysis, homology screening in related species, annotation of transcriptional elements, strucural prediction, investigation of riboswitch-like behaviour

PLOS ONE

RESEARCH ARTICLE

The Conserved *Dcw* Gene Cluster of *R. sphaeroides* Is Preceded by an Uncommonly Extended 5' Leader Featuring the sRNA UpsM

Lennart Weber[1], Clemens Thoelken[2], Marcel Volk[1], Bernhard Remes[1], Marcus Lechner[2], Gabriele Klug[1]*

1 Institute of Microbiology and Molecular Biology, IFZ, Justus-Liebig-University Giessen, Giessen, Germany, **2** Institute of Pharmaceutical Chemistry, Philipps-University Marburg, Marburg, Germany

* Gabriele.Klug@mikro.bio.uni-giessen.de

Citation: Weber L, Thoelken C, Volk M, Remes B, Lechner M, Klug G (2016) The Conserved *Dcw* Gene Cluster of *R. sphaeroides* Is Preceded by an Uncommonly Extended 5' Leader Featuring the sRNA UpsM. PLoS ONE 11(11): e0165694. doi:10.1371/journal.pone.0165694

Editor: Roy Martin Roop, II, East Carolina University Brody School of Medicine, UNITED STATES

Received: June 16, 2016

Accepted: October 17, 2016

Published: November 1, 2016

Data Availability Statement: All relevant data and accession number(s) are within the paper and its Supporting Information files. Raw files of transcriptome data are accessible via Gene Expression Omnibus (accession number GSE71844) and NCBI BioProject (accession number PRJNA343088).

Funding: This work was supported by the Deutsche Forschungsgemeinschaft (http://www.dfg.de/) (DFG Kl563/20) and the DFG funded International Research Training Group (GRK1384)

Abstract

Cell division and cell wall synthesis mechanisms are similarly conserved among bacteria. Consequently some bacterial species have comparable sets of genes organized in the *dcw (division and cell wall)* gene cluster. *Dcw* genes, their regulation and their relative order within the cluster are outstandingly conserved among rod shaped and gram negative bacteria to ensure an efficient coordination of growth and division. A well studied representative is the *dcw* gene cluster of *E. coli*. The first promoter of the gene cluster (mraZ1p) gives rise to polycistronic transcripts containing a 38 nt long 5' UTR followed by the first gene *mraZ*. Despite reported conservation we present evidence for a much longer 5' UTR in the gram negative and rod shaped bacterium *Rhodobacter sphaeroides* and in the family of *Rhodobacteraceae*. This extended 268 nt long 5' UTR comprises a Rho independent terminator, which in case of termination gives rise to a non-coding RNA (UpsM). This sRNA is conditionally cleaved by RNase E under stress conditions in an Hfq- and very likely target mRNA-dependent manner, implying its function in *trans*. These results raise the question for the regulatory function of this extended 5' UTR. It might represent the rarely described case of a *trans* acting sRNA derived from a riboswitch with exclusive presence in the family of *Rhodobacteraceae*.

Introduction

There are only rare cases for highly conserved gene clusters throughout bacterial genomes due to evolutionary dynamics. Examples for such clusters are genes for ribosomal proteins, the *atp* operon or the *dcw (division and cell wall)* gene cluster [1]. However, conservation of *dcw* genes, their regulation and especially their arrangement within the cluster are outstandingly conserved within bacterial groups of similar taxon and cell shape [2]. Besides regulatory mechanisms the conserved order of the genes may ensure an efficient coordination of growth and

(LW). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

division as assumed by the *genomic channeling hypothesis* [3]. A well described example for such conservation is the *dcw* gene cluster of gram negative and rod shaped bacteria.

Dcw gene regulation was studied intensively in *E coli*, but is not fully understood due to numerous regulatory features like internal promotors, transcript stabilities and protein ratios. It is well known that the first promoter (*mraZ*1*p*) of the gene cluster in *E. coli* (16 genes in total) gives rise to polycistronic transcripts containing a 38nt long 5' UTR followed by the first gene *mraZ* [4]. Downstream of *mraZ* transcription can potentially continue up to the last gene of the locus (*envA*) that harbors a Rho independent terminator [5].

Here we present evidence for a much longer 5' UTR in the gram negative and rod shaped bacterium *Rhodobacter sphaeroides* also present in other members of *Rhodobacteraceae*. In *R. sphaeroides* this 268 nt long 5' UTR features a Rho independent terminator 84 nt upstream of *mraZ*, which in case of transcriptional termination gives rise to a non-coding RNA of 206 nt length. This transcript was described as an orphan sRNA named RSs0682 [6], henceforth renamed UpsM (upstream sRNA of *mraZ*). Here we also show that conditional processing of UpsM requires the RNA chaperon Hfq, the endoribonuclease RNase E and induction of the RpoHI/II regulon.

Our results raise the question for the complex regulatory function of this extended 5' UTR, which might represent a rare dual function riboswitch exclusively present in the family of *Rhodobacteraceae*.

Results

dRNA-seq Hints at an Extended 5' UTR of the *Dcw* Gene Cluster of *R. sphaeroides*

The sRNA UpsM (previously RSs0682) was described as the most abundant orphan sRNA of *R. sphaeroides*, which is encoded in the intergenic region (IGR) upstream of *mraZ*, the first gene of the *dcw* (division and cell wall) gene cluster. A Rho-independent terminator was predicted at the 3' end of the sRNA locus. Identification of UpsM was based on deep sequencing of cDNA libraries using 454 pyrosequencing. The coverage of UpsM was comparably low (~2,000 reads per library) in the initial sequencing study [6]. Since sequencing technologies have rapidly evolved and nowadays generate millions of reads concomitant with visualisation of transcripts with low abundance, we re-analysed the UpsM locus in a dRNA-seq dataset that has been generated using Illumina sequencing technology (Fig 1 and S1 Fig) without and with prior TEX (terminator-5´phosphate dependent exonuclease) treatment of the RNA to enrich primary transcripts. This confirmed the presence and high abundance of UpsM and the transcriptional start site (TSS). In contrast to the low-coverage 454 pyrosequencing study, additional observations were possible. First, the processing site within UpsM, which was already detected by Northern blot analysis and 5' RACE [6], becomes apparent by a sudden decrease of reads especially after TEX treatment. Secondly, the downstream gene *mraZ* is not preceded by a separate TSS. This is surprising since *mraZ* is the first gene of the *dcw* gene cluster. Therefore it should be expressed in exponentially growing cells in the course of cell wall synthesis and cell division as described for other bacteria [4, 7–11]. This observation led to the following assumptions: *MraZ* transcription depends on the UpsM promoter and there is no additional promoter/TSS exclusively present for *mraZ*. If this is true, the terminator of UpsM has to allow read-throughs in order to guarantee transcription of *mraZ*. This has several consequences: 1.) *mraZ* or the *dcw* gene cluster have an uncommonly long 5' UTR of 268 nt in length, which has not been reported for other bacterial *dcw* clusters, which show high conservation among rod shaped and gram negative bacteria. 2.) UpsM is not an sRNA derived from an IGR (intergenic region), but rather an sRNA which is generated by transcription termination within the 5'

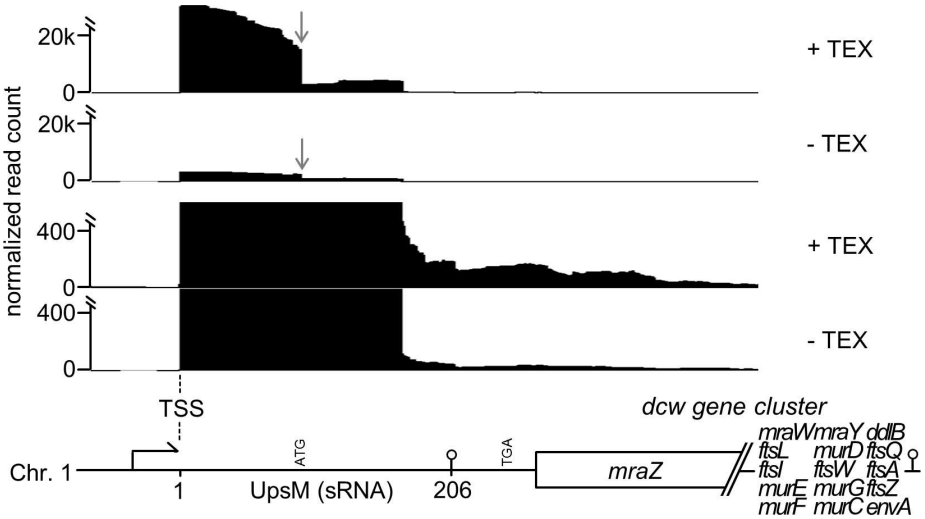


Fig 1. dRNA-seq shows a long 5' UTR of the *mraZ* gene in *R. sphaeroides*. Modified screenshots taken from IGB (integrated genome browser) visualizing the coverage at the genetic locus of *mraZ*. Shown are normalized cDNA reads on a large scale (upper two panels) and a smaller scale (lower two panels) obtained from TEX treated and untreated total RNA isolated from an exponentially and microaerobically grown *R. sphaeroides* 2.4.1 culture. The genetic context is displayed at the bottom. *mraZ* is the first gene of the *dcw* gene cluster. Position 1 reflects the TSS of sRNA UpsM (206 nt) 268 nucleotides upstream of *mraZ*. The terminator of UpsM is indicated as hairpin structure and a processing site within the sRNA is highlighted by an arrow.

doi:10.1371/journal.pone.0165694.g001

UTR of *mraZ*. 3.) In case the terminator allows read-through, the 5' UTR contains a start codon upstream and a stop codon in frame downstream of the terminator and therefore might encode a leader peptide here designated as sORF (small open reading frame).

UpsM Is a 5' UTR Derived sRNA, Which Is Conditionally Cleaved during Stress Conditions in an RNase E, Hfq and RpoHI/RpoHII-Dependent Manner

UpsM was originally identified as an sRNA processed upon $^1O_2$ stress in *Rhodobacter sphaeroides* and the RNA chaperone Hfq was shown to be required for this processing [6]. UpsM is the most abundant sRNA in *R. sphaeroides* and represents about 60% of all Hfq bound sRNAs [12]. In addition to the strong interaction of UpsM with Hfq we observed a negative growth effect under anaerobic and aerobic conditions (S2B Fig) and changes in the transcriptome in an initial microarray analysis (S1 Table) for a strain overexpressing UpsM (S2A Fig). Therefore we conclude that UpsM is functional as a *trans* acting sRNA.

To further prove that the stable 130 nt UpsM 3' fragment is generated by processing, total RNA from *R. sphaeroides* isolated 90 min after addition of methylene blue in the light to generate $^1O_2$ was treated with TEX (terminator-5´phosphate dependent exonuclease), which degrades RNAs with a monophosphate at the 5´ end but not primary transcripts that carry a 5´ triphosphate. The TEX treated RNA was compared to untreated RNA on a Northern blot (Fig 2A). The lack of the 130 nt band after TEX treatment strongly supports the assumption that this fragment is a processing product.

In gram-negative bacteria the endoribonuclease RNase E has a major role in initiation of mRNA decay [13, 14] and was also shown to be involved in the generation or processing of several sRNAs [15–18]. To test the involvement of RNAse E in UpsM processing we constructed

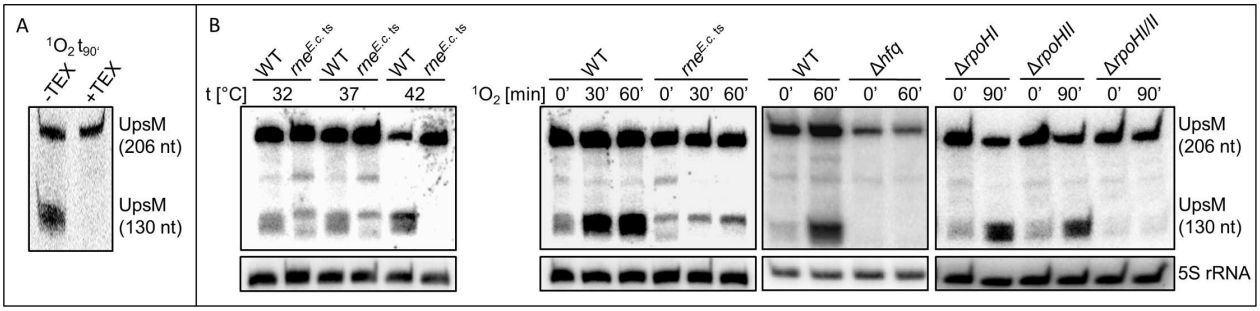


Fig 2. Northern blot analysis reveals Hfq and target mRNA dependent processing of UpsM by RNase E. (A) Detection of UpsM (206 nt) and the UpsM processing product (130 nt) by Northern blot analysis of TEX treated and untreated total RNA isolated from *R. sphaeroides* after 90 min $^1O_2$ stress. (B) Left: Comparison of the UpsM processing pattern via Northern blot analysis of total RNA isolated from *R. sphaeroides* 2.4.1 (WT) to RNA isolated from a mutant strain expressing a thermosensitive RNase E variant from *E. coli* (*rne*$^{E.c.ts}$) after growth at 32, 37 or 42˚C for 30 min or during $^1O_2$ stress. Right: Comparison of the processing pattern via Northern blot analysis to strains lacking Hfq, RpoHI, RpoHII or RpoHI and RpoHII after 0 and 60 or 0 and 90 min of $^1O_2$ stress. Signals of 5S rRNA serve as loading control.

doi:10.1371/journal.pone.0165694.g002

an *R. sphaeroides* strain with impaired RNase E activity. The endogenous *rne* gene (RSP_2131) was replaced by the *rne* gene from *E. coli* N3431 (46% blastp identity) [19, 20], which produces a temperature-sensitive RNase E due to a point mutation. As seen in Fig 2B processing of UpsM in strain *R. spharoides rne*$^{E.c.ts}$ is already impaired at 32°C or 37°C, indicating that the *E. coli* enzyme is less active than the endogenous enzyme of *R. sphaeroides*. When cells are shifted to 42°C no UpsM processing occurs in the *R. sphaeroides* strain expressing the temperature-sensitive RNase E variant. When methylene blue was added to the cultures in the light growing at 32°C accumulation of the 130 nt UpsM fragment was only weak in the *rne* mutant strain, while a strong accumulation was observed in the wild type. As shown previously a lack of Hfq also abolished UpsM processing (Fig 2B) [6].

Deletion of the endonuclease RNase III or the 3' to 5' exonuclease RNase J does not lead to an altered processing (S3A Fig), showing that those nucleases are not involved. These results demonstrate that processing of UpsM is catalyzed by RNase E. This raises the question, why RNase E dependent processing occurs only under certain growth conditions. It is highly unlikely that all conditions leading to UpsM processing go along with increased RNase E levels, increased RNase E activity or structural changes in UpsM that promote cleavage by RNase E. It is known that sRNAs are often processed together with their target mRNA [15–18]. The Hfq protein can favor sRNA-mRNA interaction and the subsequent processing [15, 21]. Thus, it is conceivable that induction of the UpsM target(s) is responsible for appearance of the 130 nt processing product under the given conditions. To date (the) target mRNA(s) of UpsM are not identified. To better define the conditions, which promote UpsM processing, we tested the effect of further stress factors and growth conditions on the UpsM pattern (S3B and S3C Fig). Neither SDS, nor ethanol, hydrogen peroxide, tBOOH nor superoxide induced processing of the UpsM transcript (S3B Fig). High NaCl concentrations induced slight processing, while $CdCl_2$ had a similar effect on processing as $^1O_2$. Heat shock resulted in very fast UpsM processing (S3C Fig). The processing product was also clearly visible in RNA isolated from stationary phase cultures.

In *R. sphaeroides* the alternative sigma factors RpoHI and RpoHII stimulate many genes in response to stress conditions including $^1O_2$, $CdCl_2$, heat and stationary phase [22–26]. Northern blots revealed that normal UpsM processing in presence of $^1O_2$ occurs in mutants either lacking RpoHI or RpoHII. However, a mutant lacking both sigma factors fails to process UpsM even under $^1O_2$ stress (Fig 2B). We conclude that interaction with a target RNA very likely

promotes UpsM cleavage by RNase E, whereas the target RNA is transcribed from a promoter, which is recognized by RpoHI as well as by RpoHII. It is known that the regulons of these two sigma factors indeed overlap [23]. No genes for known RNases are part of the RpoHI/RpoHII regulon [23, 24, 27] and UpsM heterologously expressed in *E. coli* does not show any induced processing even under stress conditions (data not shown) supporting the assumption of target-dependent processing.

The *Dcw* Gene Cluster Features a Long 5' UTR

We performed reporter assays, 5' RACE (*rapid amplification of cDNA ends*) and RT-PCRs to address the question whether *mraZ* transcription is exclusively dependent on the UpsM promoter and whether the UpsM terminator allows read-throughs in order to guarantee transcription of *mraZ*. Conversely, this means that no additional promoter is localised between the UpsM promoter and *mraZ*.

To test this we performed β-galactosidase activity assays by using reporter plasmids with *mraZ*::*lacZ* translational fusion and *mraZ* upstream regions of varying length in *R. sphaeroides* (Fig 3A). Strong activity of 130 Miller units was observed with plasmid pPHU*mraZ*UpsM containing the UpsM promoter. However, all shortened upstream regions of *mraZ* (188 nt or 67 nt) led to β-galactosidase activities similar to that observed for the empty vector control (pPHU235) proving the absence of any additional promoter closer to *mraZ*.

To further support this assumption we conducted a 5' RACE to determine 5' ends of *mraZ* mRNAs, whereby cDNA synthesis was enabled by a primer (pUpsM_mraZ_B) binding within the coding region of the *mraZ* gene. Further amplification of cDNA was done with a primer located upstream (pUpsM_B). The resulting PCR products are visible as one clear band on a gel and show the migration behaviour of fragments with 194 bp length which corresponds to the 5'end of UpsM (Fig 3B). The amplified DNA fragments were also subcloned into the pDrive vector without any further purification to determine the 5' ends precisely. The 5' end of UpsM was found in five of ten sequences, whereas the other 5' ends were distributed randomly probably due to technical reasons (Fig 3B). This experiment confirms, 1) that *mraZ* transcription depends on the promoter of UpsM, 2) the terminator of UpsM allows read-throughs leading to *dcw* transcription and 3) the 5' UTR of *mraZ* respectively *dcw* mRNAs is not processed like UpsM, since we used RNA from cells after 90 min $^1O_2$ stress and did not detect the 5' end of UpsM (130nt).

Furthermore we demonstrated read-throughs at the UpsM terminator and estimated the frequency of such events by RT-PCRs and an unconventional qRT-PCR approach with DNA free RNA from unstressed exponentially grown cultures. The two primer pairs pUpsM_A/B and pmraZ_A/B specifically amplify an UpsM or *mraZ* segment respectively (155 bp and 153 bp), whereas primer pair pUpsM_mraZ_A/B amplifies a segment (143 bp) spanning from UpsM to *mraZ* and therefore detects the read-through. RT-PCR products for all primer pairs are clearly visible on a gel, whereas none of those fragments emerge in control samples without prior reverse transcription (Fig 3C). For a rough estimate of the read-through frequency we compared amplification cycles or rather Cq values in qRT-PCR between all primer pairs in the same RNA samples instead of between different RNA samples with the very same primer pair. Since product sizes are very similar the fluorescent dye will intercalate comparably into *de novo* DNA. The mRNA level of the segment representing the read-through seems to have the lowest abundance. However, this is probably due to a terminator mediated bias and the smaller product size. In Fig 3C the relative transcript levels calculated in comparison to 16S rRNA levels are shown in relation to the transcript level detected by primer pair pUpsM_mraZ. Approximately 700 and 7 times higher RNA levels corresponding to UpsM and *mraZ* respectively were

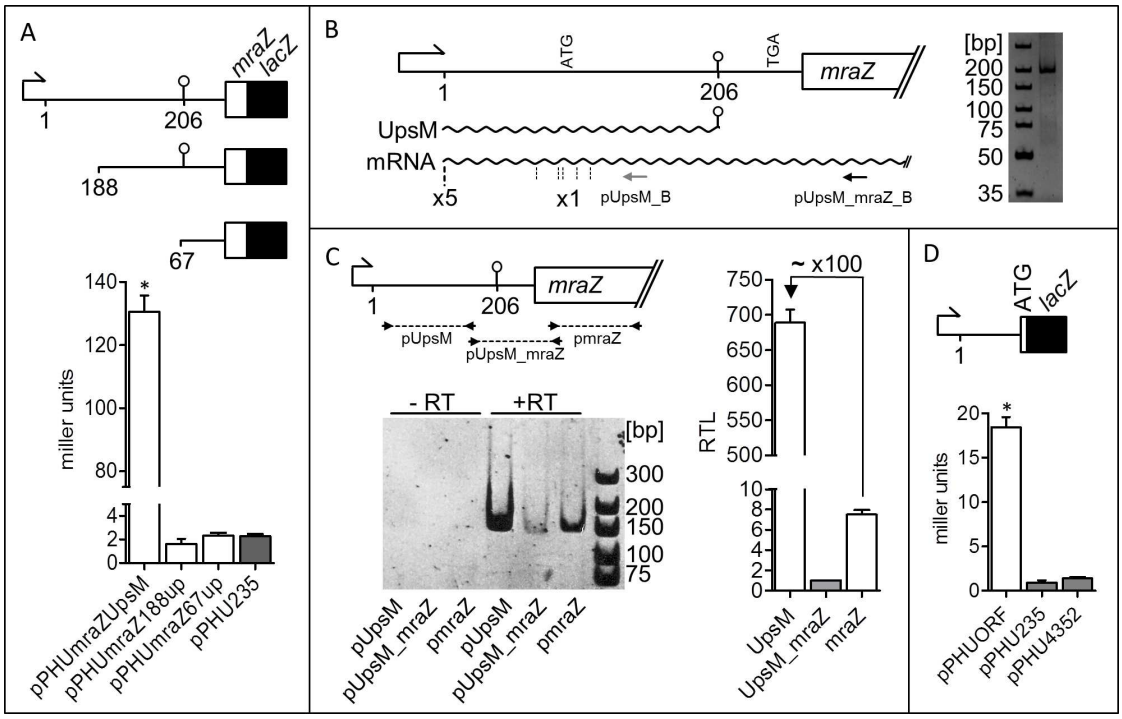


Fig 3. Transcription of *mraZ* is enabled by the UpsM promotor. (A) β-galactosidase activity assays of *R. sphaeroides* 2.4.1 with reporter plasmids with *mraZ*::*lacZ* translational fusion and *mraZ* upstream regions of varying length (long upstream region including the promotor of UpsM, 188 and 67 upstream nucleotides). pPHU235 represents the empty vector control. For each strain, three independent biological experiments with technical duplicates were performed. Error bars indicate standard deviations and an asterisk a significance level of P<0.01 compared to pPHU235. (B) 5' RACE with RNA from *R. sphaeroides* 2.4.1 after 90 min of $^1O_2$ stress. cDNA was generated with the primer depicted as black arrow, whereas cDNAs were amplified by the primer shown as grey arrow. The PCR product was visualized on a gel (10% PAA/TBE) by ethidium bromide staining. 5' ends (dashed lines) identified by subcloning and sequencing and their corresponding frequencies are highlighted. (C) qRT-PCR products of primer pairs pUpsM, pmraZ (155 bp and 153 bp, both specific for the corresponding mRNA segments) and pUpsM_*mraZ* (143 bp, spanning from UpsM to *mraZ*) visualized on a gel (10% PAA/TBE) by ethidium bromide staining. Samples without initial RT step were loaded as control. On the right relative transcript levels are shown in relation to the product quantity of primer pair pUpsM_*mraZ*. qRT-PCRs were performed in technical duplicates with RNA from three biological independent and unstressed *R. sphaeroides* 2.4.1 cultures. Error bars indicate standard deviations. (D) β-galactosidase activity assays of *R. sphaeroides* 2.4.1 conjugated with a reporter plasmid with translational *lacZ* fusion to the start codon (ATG) within the UpsM gene in comparison to the promoter-less empty vector control (pPHU235) and a control plasmid (pPHU4352) containing a strong 16S rRNA promoter. For each strain, three independent biological experiments with technical duplicates were performed. Error bars indicate standard deviations and an asterisk a significance level of P<0.01 compared to both controls.

doi:10.1371/journal.pone.0165694.g003

detected. As demonstrated *mraZ* transcripts have long 5' UTRs and contain the UpsM locus. Therefore part of the DNA amplified by the primer pair pUpsM originates from full-length transcripts also encoding *mraZ*. In other words, *mraZ* transcripts are amplified not only by the primer pair pmraZ but also by pUpsM primers. By taking this into consideration we estimate that one read-through event or rather *mraZ* transcription takes place about once in 100 transcription events under the given experimental conditions.

We were able to show convincingly that *mraZ* has a long 5' UTR under transcriptional control of the UpsM promoter. In addition the 5' UTR contains a start codon upstream and a stop codon in frame downstream of the terminator of UpsM and therefore might encode a leader peptide here designated as sORF (small open reading frame). To test translation initiating at this start codon we performed β-galactosidase activity assays of a reporter plasmid with an

sORF::*lacZ* translational fusion containing the upstream region including P_{UpsM} in *R. sphaeroides*. The sORF::*lacZ* fusion on plasmid pPHUORF resulted in low but significant higher β-galactosidase activities of approximately 18 Miller units in comparison to the promoter-less empty vector control pPHU235 and control plasmid pPHU4352 containing a strong 16S rRNA promoter (Fig 3D). This indicates translation of the ORF, but final proof by a direct detection of the hypothetical peptide is missing.

The 5' UTR of the *Dcw* Gene Cluster in *Rhodobacteraceae* Differs from Other Bacteria

The upstream region of the *mraZ* gene was compared to that of other species. Using public available deep sequencing data obtained from the NCBI SRA database [28], we annotated TSS (transcription start sites) within 300 nt upstream of the gene locus. Reads were mapped to the respective genomes using segemehl [29] with default parameters after quality trimming using Trimmomatic [30] at a quality threshold of 25 in a sliding window of size 3. Only reads of size >14 nt were considered. See S4 Fig for details. Rho-independent Terminators in the 5'-region were predicted using TransTermHP [31]. The highest scoring hit was assumed to be the terminator whenever its MFE (minimum free energy) was below -10 kcal/mol according to RNAfold [32].

A summary is shown in Fig 4. *Rhodobacteraceae* consistently show a long 5' UTR with a strong terminator and no additional TSS close to *mraZ* in a particularly striking manner. In

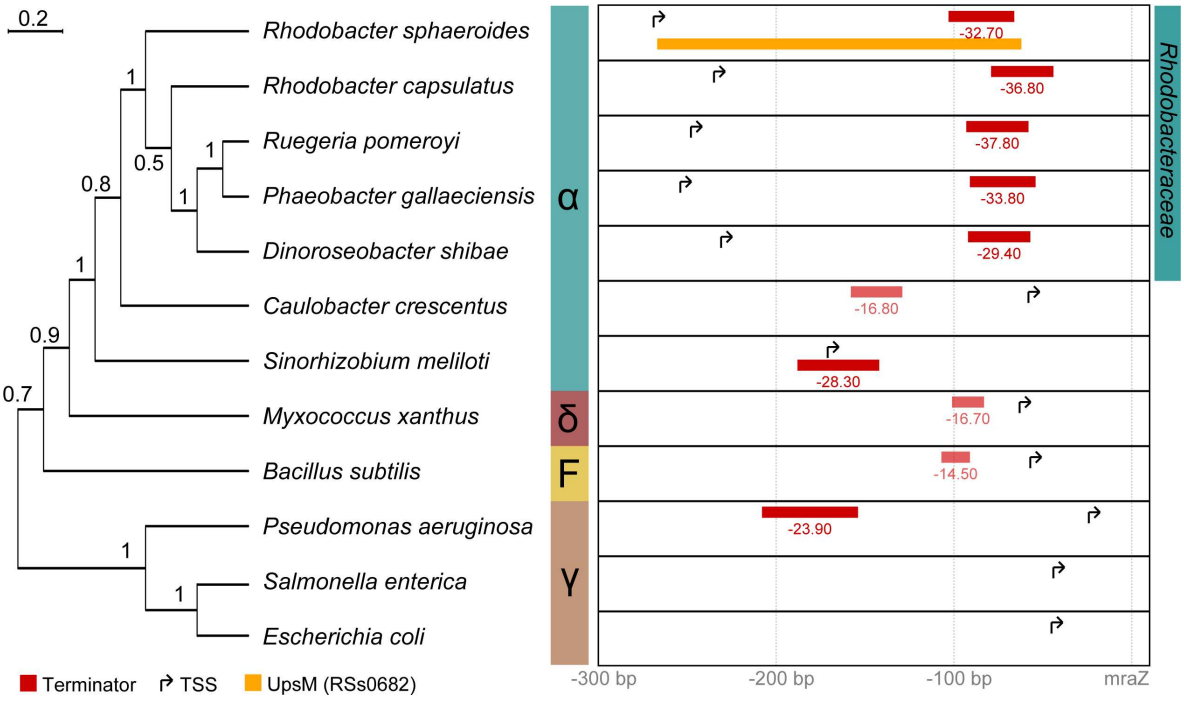


Fig 4. Comprehensive view of *mraZ* upstream regions in different species. Terminator predictions are indicated in red. Respective energies are given in kcal/mol. Regions between Start and Stop codons in frame are shown as grey bars. Transcription start sites are derived from public available deep sequencing data (see S4 Fig for details). The phylogenetic tree was build using clustalx [58] (NJ, 10000 bootstraps) based on a clustalOmega [59] alignment of the respective *mraZ* coding regions. Bootstrap support values are indicated. Seemingly the long *mraZ* 5' UTR with an intrinsic terminator is special to the family of *Rhodobacteraceae*.

doi:10.1371/journal.pone.0165694.g004

contrast, other species apart from *Rhodobacteraceae* typically have an exclusive TSS closer to *mraZ* resulting in 5' URTs of $\leq$ 70 nt, including the closely related Alphaproteobacterium *Caulobacter crescentus*. Moreover upstream terminators are not found at all or are predicted to be rather weak in those species. A notable exception is *Sinorhizobium meliloti* by exhibiting 5' UTR mediocre in length (170 nt) whereby the TSS is located within a predicted terminator site. However, whether this upstream TSS contributes to the *mraZ* expression as shown for *R. sphaeroides* remains speculative but has never been reported. Taken together our data suggests that a long *mraZ* 5' UTR with intrinsic terminator generating an sRNA combined with no separate TSS for *mraZ* is an exclusive feature of the family of *Rhodobacteraceae*.

Predicted Secondary Structure and Folding Landscape of UpsM

The secondary structure prediction of UpsM was evaluated *in silico* using RNAfold [32]. It consists of four structured regions (R1-4). R1 is located at the 5' end. It consists of several short hairpins isolated from the remaining RNA by a ~11 nt long stem followed by an unpaired region of 10 nt. R2 is a long hairpin with a small bulge, R3 a short hairpin with a stem of only 4 nt. R4 is located at the 3' end of the molecule and corresponds to the predicted, strong terminator structure (Fig 5A).

The secondary structure elements R2/3 and R4 show similarities to typical riboswitches (R2/3 aptamer region, R4 terminator) [33]. Hence we constrained the fold such that the terminator would not form. Strikingly, the refolding event leads to an interaction of R2 and R4 owing a loss of ~20% energy (15.9 kcal/mol) (Fig 5B). Based on this analysis, the *dcw* 5' UTR has the potential to represent a riboswitch.

To verify our assumption, we aligned the 5'-UTR regions of all *Rhodobacteraceae* in our dataset (cropped at the predicted terminator) using mlocarna [34] which respects sequence

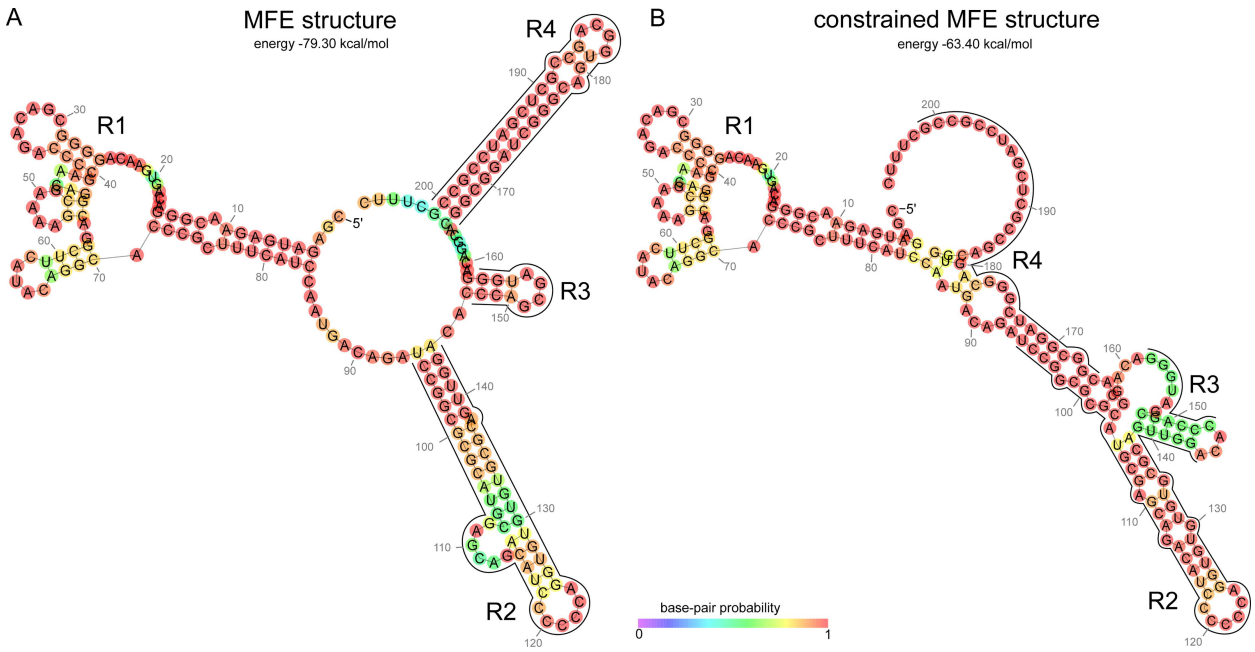


Fig 5. Structural analysis of UpsM. Analogous structured regions are indicated as R1-R4. RNAfold structure of UpsM in *R. sphaeroides* with and without constraint terminator (R4).

doi:10.1371/journal.pone.0165694.g005

and secondary structure at the same time. The alignment was folded using RNAalifold from the Vienna RNA 2.0 package [32] with no constraints and, analogously to above, constraint such that the terminator would not be allowed to form. S5A Fig shows the resulting structures with conservation. S5B Fig indicates the underlying alignment with secondary structure indications.

The consensus structure implies that R1 is hardly conserved. Moreover, the predicted transcription start site is further downstream in other species, leading to shorter transcripts (e.g. 19 nt of R1 are missing in *D. shibae*). We thus assume that R1 is relevant for the function of UpsM. Other than that, the consensus structure well resembles the UpsM structure, indicating that R2-4 are conserved among *Rhodobacteraceae*. Constraint folding however does not induce an interaction of R2 and R4 in this case. Hence, the putative riboswitch function is seemingly not conserved in *Rhodobacteraceae* but, if so, an evolutionary trait special to *R. sphaeroides*.

The folding landscape of UpsM was predicted using RNAsubopt, RNAfold and barriers from the Vienna RNA 2.0 package [32] as well as treekin [35]. A summary is shown in S6 Fig. The four most probable structures of UpsM mainly differ in the terminator hairpin (R4, states 1, 4 and 12) while one variant slightly differs in R1. States 1 and 4 represent the most pronounced folding minima with the highest probability (S6A and S6B Fig). Potential interactions between the structured regions are not observed. These findings indicate a rather stable structure that requires an additional partner (e.g RNA, protein or ligand) for substantial refold.

Discussion

In this study we further characterize the sRNA UpsM (previously RSs0682) and demonstrate that it is derived from the 5´ UTR of the mRNA for *mraZ*, the first gene of the *dcw* gene cluster. Overexpression of UpsM leads to a mild growth defect and to a change in the global gene expression pattern as well under aerobic conditions as under photooxidative stress conditions, proving that UpsM is also functional in *trans*. Further experiments need to clarify whether altered mRNA levels are due to direct base pairing or rather to binding of high amounts of Hfq. Analysis of a dRNA-seq dataset of high coverage and Northern blot analysis of TEX treated RNA from $^1O_2$ stressed cultures confirmed a processing step from UpsM (206nt) to UpsM (130nt). The UpsM processing pattern under different stress conditions and in various mutant strains showed that processing requires the RNA chaperon Hfq, the endoribonuclease RNase E and the alternative sigma factors RpoHI/II. mRNAs being part of the RpoHI/II controlled stress regulon are induced similar to the processing of UpsM upon $^1O_2$ stress [6, 12, 23], heat stress [24] and in stationary phase (unpublished). Therefore we assume that UpsM is bound and stabilized by Hfq prior to target recognition. However, under stress conditions a target mRNA is expressed which may form a duplex with UpsM mediated by Hfq. In the course of base pairing the structure of UpsM might be altered and becomes susceptible to RNase E cleavage. A similar mechanism was first described for the Hfq dependent sRNA RyhB, which is degraded in an RNase E dependent manner upon binding to the target mRNA *sodB* in *E. coli* [15]. Moreover, an interaction of Hfq with the scaffolding domain of RNase E for the purpose of a recruitment of RNase E to sRNA-mRNA hybrids has been discussed [36] and might also be true for *R. sphaeroides*.

The dRNA-seq data of high coverage based on exponentially growing *R. sphaeroides* cultures, did not show a TSS exclusive for *mraZ*, despite the fact that *mraZ* represents the first gene of the *dcw* gene cluster. Therefore and based on dRNA-seq data we hypothesized that transcription of *mraZ* depends on the UpsM promoter, which implicates that the terminator of UpsM allows read-throughs in order to guarantee transcription of *mraZ* with a long 5' UTR of 268 nt in length. We were able to proof this assumption by 5' RACE and reporter assays with

different fragments of the *mraZ* 5′ UTR and estimated the read-through at the UpsM terminator to takes place once in 100 transcription events under our experimental conditions. This might ensure sufficient proximal *dcw* gene transcription, since UpsM is the most frequently transcribed sRNA in *R. sphaeroides* [6]. Taken together our data demonstrates that *mraZ* or polycistronic *dcw* gene mRNAs of *R. sphaeroides* feature long 5' UTRs, with the consequence that UpsM has to be reckoned as a 5' UTR derived sRNA rather than an orphan sRNA derived from an IGR as previously assumed [6].

In recent years transcriptome wide identification of sRNAs revealed that apart from intergenic regions especially 3' UTRs serve as a reservoir for sRNAs being part in the Hfq network [37, 38] and their functions in *trans* was reported [39–41]. In comparison, 5' UTR- or riboswitch-derived sRNAs have been described only occasionally and in general without clear functional assignment in *trans*. A handful of putative sRNAs resulting from processed 5' UTRs are mentioned in a transcriptome study of *Yersinia pseudotuberculosis* [42]. For *E. coli* prematurely terminated transcripts from 5' leader sequences of *ybjM*, *ynaE*, *ydfK*, *mdtJ*, *typA*, *yhiI*, and *dinQ* were detected [43], among which *ynaE*, *ydfK* and *ybjM* 5' leader fragments co-immunoprecipitate with Hfq [44]. In the same study an sRNA corresponding to a 5' UTR segment of the *adhE* mRNA was identified, possibly generated by RNase III [44]. A cloning based screen for ncRNAs in *E. coli* led to the discovery of a few 5' UTR derived RNA fragments, which correspond to riboswitches or to be precise L-box, THI box and RFN box elements known to be required for the regulation of mRNA or protein synthesis by attenuation or RBS accessibility [45]. However, neither an association to Hfq nor any other proof for a function of those RNA fragments in *trans* was provided in this study. Recently 14 sRNAs which also might serve as 5' UTRs were predicted in *Vibrio cholera*, but only for Vcr043 an Hfq dependent stability was shown indicating a function in *trans* [46]. To our knowledge and surprisingly only once dual function 5' UTRs were described so far, which act as riboswitch and generate sRNAs in the course of attenuation with regulatory function in *trans*. This was described for SreA and SreB, two S-adenosylmethionine (SAM) riboswitches in *Listeria monocytogenes*. SreA and SreB are transcribed together with downstream genes encoding proteins involved in methionine and cysteine transport or metabolism. Upon binding of the ligand SAM transcription is attenuated causing expression of SreA and SreB as short transcripts, which control expression of the virulence regulator PrfA precisely as sRNAs by binding to the 5‘ region of its mRNA [47]. The authors assumed SreA/SreB to be just the first example of a novel distinct class of riboswitch derived sRNAs, without expecting that no further example was reported over the past seven years. In this study we were able to describe a similar example, since UpsM can be transcribed with the downstream gene, but is also generated as stable sRNA by an intrinsic terminator in the 5' leader. However, we were only able to provide evidence for a function of UpsM in *trans* by strong Hfq dependency and target dependent cleavage of the sRNA, whereas we cannot present any experimental evidence for a riboswitch at the intrinsic terminator in the leader sequence so far. Its prominent position in the 5' UTR of the *dcw* gene cluster and combined with the potential breakup the terminator structure R4 due to an interaction with R2, however, makes UpsM a reasonable candidate for a riboswitch. Regulatory features and a potential function as riboswitch will be subject to future investigation.

The regulatory features of the *mraZ/dcw* 5' UTR might be even more complex, since we have indications for weak translational activity at an sORF overlapping with the terminator of UpsM. *Cis*-regulatory elements as such sORFs sometimes also termed leader peptides or uORFs (upstream open reading frame) in eukaryotes are often involved in transcriptional or translation attenuation [48, 49]. However, this sORF is only present in *R. sphaeroides* species and presence of the resulting peptide is not unequivocally proven.

The *mraZ/dcw* 5' UTR or rather UpsM may represent not only the second example for a class of riboswitch derived sRNAs, but also is the first extended 5' UTR described for the *dcw* gene cluster, despite outstanding conservation among rod shaped and gram negative *bacteria* [1–3]. In *E. coli* the first promoter (*mraZ1p*) of the gene cluster (16 genes in total) gives rise to polycistronic transcripts containing a short 38nt long 5' UTR followed by the first gene *mraZ* [4, 7]. However, in this study we have been able to describe a much longer 5' UTR of 268 nt in length featuring *trans*-regulatory and potential *cis*-regulatory elements (summarized in S7 Fig). Therefore we were interested whether we would find similar 5' leaders in the same genetic context of other bacteria. Our results suggest that long 5'UTRs with intrinsic terminators are exclusively present in members of *Rhodobacteraceae*.

Material and Methods

Bacterial Strains and Growth Conditions

Bacterial strains used in this study are listed in S2 Table. Details on their construction are given in S1 File. *R. sphaeroides* strains were cultivated at 32°C in malate minimal-salt medium [50]. To grow the cells aerobically, cultures were either gassed with air in Meplat bottles to attain a concentration of 160 to 180 μM of dissolved oxygen or by continuous shaking of Erlenmeyer flasks containing 20% culture by volume at 140 rpm. For microaerobic growth conditions, having a dissolved oxygen concentration of about 25 μM, Erlenmeyer flasks containing 80% culture by volume were shaken at 140 rpm. For anaerobic growth in the dark in the presence of 60 mM DMSO as electron acceptor we used completely filled screw-cap Meplat bottles, completely filled and sealed with Parafilm. When necessary tetracycline (2 μg ml^{-1}), kanamycin (25 μg ml^{-1}) or spectinomycin (10 μg ml^{-1}) was added to liquid and solid growth media (1.6% agar). Photooxidative stress conditions were generated as described earlier [51], except the final concentration of methylene blue (0.2 μM) (Sigma-Aldrich; M9140). Other stress conditions were generated by a final concentration of 250 mM NaCl, 10 μM $CdCl_2$, 0.005% SDS, 2.5% ethanol, 300 μM tBOOH, 1 mM H_2O_2 and 250 μM paraquat (O_2^-) or by temperature shift to 42°C. To culture *E. coli* strains, cells were continuously shaken at 180 rpm in Luria–Bertani medium at 37°C or grown on solid growth media containing 1.6% (w/v) agar. When necessary kanamycin (25 μg ml^{-1}) or tetracycline (20 μg ml^{-1}), ampicillin (200 μg ml^{-1}) or spectinomycin (10 μg ml^{-1}) was added to the media.

Northern Blot Analysis

Northern blots were performed as described earlier [6]. Oligodeoxynucleotides used for end-labeling with [γ-^{32}P]-ATP (Hartmann Analytic; SRP-301) by T4 polynucleotide kinase (Fermentas; #EK0031) are listed in S3 Table. A low stringency Church buffer was used for hybridization [52]. Membranes were washed in 5x SCC buffer + 0.1% SDS. After exposure on phosphoimaging screens (Bio-Rad), images were analyzed by the 1D-Quantity One software (Bio-Rad).

Isolation of Total RNA

Total RNA used for Northern blot, 5' RACE and real time RT-PCR was isolated by the hot phenol method [53]. To remove remaining traces of DNA, samples were treated with 6 U of DNaseI (Invitrogen; #18047019) per 1 μg of RNA. Absence of DNA contamination was confirmed by PCR with primers targeting *gloB* (RSP_0799) (S3 Table).

5' RACE

To determine 5' mRNA ends of *mraZ* using 5' rapid amplification of cDNA ends (RACE), 3 μg of DNA free total RNA isolated from wild type cells after 90 minutes of $^1O_2$ stress were reverse transcribed into cDNA by using avian myeloblastosis virus reverse transcriptase (Promega) and gene-specific primer pUpsM_mraZ_B (S3 Table). A second amplification was done with primer pUpsM_B (S3 Table). The 5′RACE protocol was performed as described previously [24].

qRT-PCR

The One-Step Brilliant III QRT-PCR Master Mix Kit (Agilent) was used for reverse transcription and following PCR as described in the manufacturer's manual but in 10 μl volumes containing 2 μl DNA free RNA in the concentration 0.2 ng/μl. Runs in independent biological triplicates with technical duplicates were done by the use of a Bio-Rad CFX96 Real Time System. Cq values at the auto calculated RFU were extracted with the corresponding software Bio-Rad CFX Manager. mRNA levels were calculated in relation to the mRNA levels of 16S rRNA similar to Pfaffl [54], but with fixed primer efficiencies of 2.0 and a fixed denominator of 1.0, since Cq values of different primer pairs were compared and the very same RNA sample was used. Therefore the resulting formula is: Ratio = $2^{\Delta Cq(16S\ -A)}$, whereas A is Cq of primer pair pUpsM, pUpsM_marZ or pmraZ. Primers are listed in S3 Table.

β-Galactosidase Activity Assay

β-galactosidase activity was measured in conjugants obtained after transferring the respective reporter plasmids via di-parental conjugation from *E. coli* to *R. sphaeroides*. Three independent liquid cultures, inoculated with equal numbers of colonies, were grown microaerobically and diluted to OD_{660} 0.2, before reaching stationary phase. Three samples of 1 ml were collected in early exponential growth phase (OD_{660} 0.4). Measurements of β-galactosidase activity were carried out as described previously [22].

RNA Treatment, Library Preparation and Sequencing

DNA free total RNA isolated from exponential and microaerobic cultures were treated with TEX (Epicentre #TER51020) to enrich primary transcripts [55]. For *R. sphaeroides* RNA Illumina cDNA libraries were prepared by vertis Biotechnology AG, Germany (http://www.vertis-biotech.com/) as described before without prior RNA fragmentation or size fractionation [50]. Illumina cDNA libraries resulting from TEX treated *R. capsulatus* RNA were generated as described before without prior rRNA depletion [42]. cDNA libraries were sequenced on a HiSeq 2000 or HiSeq 2500 machine in single-read mode running 100 cycles. Raw files for *R. sphaeroides* have been deposited in the National Center for Biotechnology Information Gene Expression Omnibus (GEO) [56] and are accessible via the GEO accession GSE71844. Raw files for *R. capsulatus* are accessible via BioProject Accession PRJNA343088.

Microarray Analysis

Microarray analysis was performed as described before [26, 57]. Total RNA, obtained from 6 independent cultures per strain later hybridized with a duplicate of arrays was chemically labeled with Cy3 and Cy5 (Kreatech; EA-022/EA-023), respectively. Multiarray analysis was performed with the Bio-conductor package Limma for R. On the basis of calculated MA plots, genes were considered reliable if the average signal intensity [A-value: 1/2 log2 (Cy3×Cy5)] was ≥ 12. To filter out potentially insignificant changes among genes that passed the reliability

criterion, a cutoff value was applied; i.e., those genes were retained whose average expression value of the overexpression strain (a) compared with the average value of the control treatment (b) was either a $\log_2$ fold change of $\geq$ 0.65 or $\leq$ -0.65. Microarray data are deposited in the Gene Expression Omnibus (GSE87789).

Supporting Information

S1 Fig. Nucleotide sequence of the *mraZ* 5' UTR. Sequence of UpsM is shaded in grey. The corresponding terminator is underlined. The hypothetical sORF coding region starting with ATG and the *mraZ* coding region starting with GTG are depicted by bold letters.
(TIF)

S2 Fig. (A) Altered UpsM transcript level shown by Northern blot analysis of total RNA of the overexpression strain *R. sphaeroides* 2.4.1 pBBRUpsMx2 after 60 min $^1O_2$ stress in comparison to the wild-type strain harboring the empty vector (pBBR1MCS2). Signals of 5S rRNA serve as loading control. (B) Aerobic, microaerobic and anerobic growth of the overexpression strain *R. sphaeroides* 2.4.1 pBBRUpsMx2 in comparison to the wild-type strain harboring the empty vector (pBBR1MCS2). The optical density at 660 nm (OD_{660}) was determined over time, and growth is indicated as continuous line. All graphs represent the mean of three biological independent experiments. Error bars indicate the standard deviation at each time point measured.
(TIF)

S3 Fig. Processing pattern of UpsM under various ongoing stress conditions and in strains lacking RNases shown by Northern blot analysis of total RNA isolated from *R. sphaeroides*. (A) Processing pattern of UpsM in strains lacking RNase III (Δ*rnc*) or RNase J (Δ*rnj*). Signals of 5S rRNA serve as loading control. (B) Stress conditions were generated by a final concentrations of 0.005% SDS, 2.5% ethanol, 300 μM tBOOH, 1mM H_2O_2 and 250 μM paraquat (O_2^-). (C) Stress conditions were generated by a final concentrations of 0.2 μM methylene blue in the presence of 800 Wm^{-2} white light ($^1O_2$), 250 mM NaCl and 10 μM $CdCl_2$ or by stationary phase or growth under heat stress at 42°C.
(TIF)

S4 Fig. Genomic regions upstream of *mraZ* in other species. Underlying deep sequencing data was used to predict transcriptional start sites. Layout resembles Fig 4. Genome IDs are indicated at the y-axis. Sources are indicated at the right.
(TIF)

S5 Fig. Structural analysis of UpsM. Analogous structured regions are indicated as R1-R4. (A) RNAfold structure of UpsM in *R. sphaeroides* with and without constraint terminator (R4) and consensus structure of aligned sequences for all *Rhodobacteraceae* without constraint terminator (R4). (B) RNAalifold alignment with structural annotation and indicated terminator constraint (x = bases forced to be unpaired). An interaction of R2 and R4 occurs when applying terminator constraints. This is however not resembled by the consensus structure.
(TIF)

S6 Fig. Folding state analysis of UpsM. (a) Barrier tree of all suboptimal structures of the unprocessed sRNA. Four main species with favorable energies were identified (1, 4, 8, 12). (B) Population density of these structures over time (no unit). (C) Structural representations of the four most favorable states.
(TIF)

S7 Fig. Model summarizing the results of this publication. The 268 nt long 5' UTR of *mraZ*, first gene of the *dcw* (division and cell wall) gene cluster, comprises a Rho independent terminator, which in case of termination gives rise to the 206 nt long non-coding RNA UpsM (upstream sRNA *mraZ*). Under stress conditions this sRNA is conditionally cleaved by RNase E in an Hfq- and likely in a target mRNA-dependent manner, whereas the corresponding target mRNA is controlled by an RpHI/II dependent promotor.
(TIF)

S1 File. Strain construction.
(PDF)

S1 Table. Gene expression of an UpsM overexpression strain *R. sphaeroides* 2.4.1 pBBRUpsMx2 was analysed in comparison to the strain *R. sphaeroides* 2.4.1 pBBR1MCS2 harbouring the empty vector to get first insights into the biological function of UpsM. The transcriptome of both strains was compared by microarray analysis during exponential growth under aerobic and non-stress conditions and after 90 min of $^1O_2$ stress. For both conditions a biological duplicate of arrays was hybridized with RNA from three biological independent cultures per strain. A Pearson correlation coefficient between the replica of 0.97 and 0.95 was calculated. Changes in expression levels of protein-coding genes passing the selection criteria of microarray analysis, which is a reliable A-value ≥ 12 and a log2 fold change of > 0.65 or < -0.65 between the two strains, are shown.
(DOCX)

S2 Table. Strains and plasmids used in this study
(DOCX)

S3 Table. Oligonucleotides used in this study.
(DOCX)

Acknowledgments

We thank Kerstin Haberzettl for help with construction of the *rne* mutant strain and Sven Findeiss for discussion of the folding state analysis.

Author Contributions

Conceptualization: LW GK.

Data curation: GK ML.

Formal analysis: LW CT ML BR.

Funding acquisition: GK.

Investigation: LW BR MV CT ML.

Methodology: LW.

Project administration: GK LW.

Resources: GK ML.

Supervision: LW GK.

Validation: GK LW.

Visualization: LW.

Writing – original draft: LW ML GK.

Writing – review & editing: LW GK.

References

1. Tamames J. Evolution of gene order conservation in prokaryotes. Genome Biol. 2001; 2(6): RESEARCH0020. PMID: 11423009; PubMed Central PMCID: PMC33396.
2. Tamames J, Gonzalez-Moreno M, Mingorance J, Valencia A, Vicente M. Bringing gene order into bacterial shape. Trends Genet. 2001; 17(3):124–6. PMID: 11226588.
3. Mingorance J, Tamames J, Vicente M. Genomic channeling in bacterial cell division. J Mol Recognit. 2004; 17(5):481–7. doi: 10.1002/jmr.718 PMID: 15362108.
4. Vicente M, Gomez MJ, Ayala JA. Regulation of transcription of cell division genes in the *Escherichia coli* dcw cluster. Cell Mol Life Sci. 1998; 54(4):317–24. PMID: 9614967. doi: 10.1007/s000180050158
5. de la Fuente A, Palacios P, Vicente M. Transcription of the *Escherichia coli dcw* cluster: evidence for distal upstream transcripts being involved in the expression of the downstream *ftsZ* gene. Biochimie. 2001; 83(1):109–15. PMID: 11254983.
6. Berghoff BA, Glaeser J, Sharma CM, Vogel J, Klug G. Photooxidative stress-induced and abundant small RNAs in *Rhodobacter sphaeroides*. Mol Microbiol. 2009; 74(6):1497–512. doi: 10.1111/j.1365-2958.2009.06949.x PMID: 19906181.
7. Eraso JM, Markillie LM, Mitchell HD, Taylor RC, Orr G, Margolin W. The Highly Conserved MraZ Protein Is a Transcriptional Regulator in *Escherichia coli*. J Bacteriol. 2014; 196(11):2053–66. doi: 10.1128/Jb.01370-13 PMID: WOS:000335909500014.
8. Hara H, Yasuda S, Horiuchi K, Park JT. A promoter for the first nine genes of the *Escherichia coli mra* cluster of cell division and cell envelope biosynthesis genes, including *ftsI* and *ftsW*. J Bacteriol. 1997; 179(18):5802–11. PMID: 9294438; PubMed Central PMCID: PMC179470.
9. Selinger DW, Saxena RM, Cheung KJ, Church GM, Rosenow C. Global RNA half-life analysis in *Escherichia coli* reveals positional patterns of transcript degradation. Genome Res. 2003; 13(2):216–23. doi: 10.1101/gr.912603 PMID: 12566399; PubMed Central PMCID: PMC420366.
10. Mengin-Lecreulx D, Ayala J, Bouhss A, van Heijenoort J, Parquet C, Hara H. Contribution of the Pmra promoter to expression of genes in the *Escherichia coli mra* cluster of cell envelope biosynthesis and cell division genes. J Bacteriol. 1998; 180(17):4406–12. PMID: 9721276; PubMed Central PMCID: PMC107448.
11. Carrion M, Gomez MJ, Merchante-Schubert R, Dongarra S, Ayala JA. *mraW*, an essential gene at the *dcw* cluster of *Escherichia coli* codes for a cytoplasmic protein with methyltransferase activity. Biochimie. 1999; 81(8–9):879–88. PMID: 10572301.
12. Berghoff BA, Glaeser J, Sharma CM, Zobawa M, Lottspeich F, Vogel J, et al. Contribution of Hfq to photooxidative stress resistance and global regulation in *Rhodobacter sphaeroides*. Mol Microbiol. 2011; 80(6):1479–95. doi: 10.1111/j.1365-2958.2011.07658.x PMID: 21535243.
13. Belasco JG. All things must pass: contrasts and commonalities in eukaryotic and bacterial mRNA decay. Nat Rev Mol Cell Biol. 2010; 11(7):467–78. doi: 10.1038/nrm2917 PMID: 20520623; PubMed Central PMCID: PMC3145457.
14. Carpousis AJ. The RNA degradosome of *Escherichia coli*: An mRNA-degrading machine assembled on RNase E. Annu Rev Microbiol. 2007; 61:71–87. doi: 10.1146/annurev.micro.61.080706.093440 PMID: WOS:000250965600006.
15. Masse E, Escorcia FE, Gottesman S. Coupled degradation of a small regulatory RNA and its mRNA targets in *Escherichia coli*. Genes Dev. 2003; 17(19):2374–83. doi: 10.1101/gad.1127103 PMID: 12975324; PubMed Central PMCID: PMC218075.
16. Pfeiffer V, Papenfort K, Lucchini S, Hinton JC, Vogel J. Coding sequence targeting by MicC RNA reveals bacterial mRNA silencing downstream of translational initiation. Nat Struct Mol Biol. 2009; 16(8):840–6. doi: 10.1038/nsmb.1631 PMID: 19620966.
17. Prevost K, Desnoyers G, Jacques JF, Lavoie F, Masse E. Small RNA-induced mRNA degradation achieved through both translation block and activated cleavage. Genes Dev. 2011; 25(4):385–96. doi: 10.1101/gad.2001711 PMID: 21289064; PubMed Central PMCID: PMC3042161.
18. Saramago M, Barria C, Dos Santos RF, Silva IJ, Pobre V, Domingues S, et al. The role of RNases in the regulation of small RNAs. Curr Opin Microbiol. 2014; 18:105–15. doi: 10.1016/j.mib.2014.02.009 PMID: 24704578.

19. Apirion D. Isolation, genetic mapping and some characterization of a mutation in *Escherichia coli* that affects the processing of ribonuleic acid. Genetics. 1978; 90(4):659–71. PMID: 369943; PubMed Central PMCID: PMC1213911.
20. Goldblum K, Apririon D. Inactivation of the ribonucleic acid-processing enzyme ribonuclease E blocks cell division. J Bacteriol. 1981; 146(1):128–32. PMID: 6163761; PubMed Central PMCID: PMC217061.
21. Waters LS, Storz G. Regulatory RNAs in bacteria. Cell. 2009; 136(4):615–28. doi: 10.1016/j.cell.2009.01.043 PMID: 19239884; PubMed Central PMCID: PMC3132550.
22. Nuss AM, Adnan F, Weber L, Berghoff BA, Glaeser J, Klug G. DegS and RseP homologous proteases are involved in singlet oxygen dependent activation of RpoE in *Rhodobacter sphaeroides*. PLoS One. 2013; 8(11):e79520. doi: 10.1371/journal.pone.0079520 PMID: 24223961; PubMed Central PMCID: PMC3818230.
23. Nuss AM, Glaeser J, Berghoff BA, Klug G. Overlapping alternative sigma factor regulons in the response to singlet oxygen in *Rhodobacter sphaeroides*. J Bacteriol. 2010; 192(10):2613–23. doi: 10.1128/JB.01605-09 PMID: 20304993; PubMed Central PMCID: PMC2863570.
24. Nuss AM, Glaeser J, Klug G. RpoH(II) activates oxidative-stress defense systems and is controlled by RpoE in the singlet oxygen-dependent response in Rhodobacter sphaeroides. J Bacteriol. 2009; 191(1):220–30. doi: 10.1128/JB.00925-08 PMID: 18978062; PubMed Central PMCID: PMC2612413.
25. Billenkamp F, Peng T, Berghoff BA, Klug G. A cluster of four homologous small RNAs modulates C1 metabolism and the pyruvate dehydrogenase complex in *Rhodobacter sphaeroides* under various stress conditions. J Bacteriol. 2015; 197(10):1839–52. doi: 10.1128/JB.02475-14 PMID: 25777678; PubMed Central PMCID: PMC4402390.
26. Adnan F, Weber L, Klug G. The sRNA SorY confers resistance during photooxidative stress by affecting a metabolite transporter in *Rhodobacter sphaeroides*. RNA Biol. 2015; 12(5):569–77. doi: 10.1080/15476286.2015.1031948 PMID: 25833751; PubMed Central PMCID: PMC4615379.
27. Dufour YS, Imam S, Koo BM, Green HA, Donohue TJ. Convergence of the transcriptional responses to heat shock and singlet oxygen stresses. PLoS Genet. 2012; 8(9):e1002929. doi: 10.1371/journal.pgen.1002929 PMID: 23028346; PubMed Central PMCID: PMC3441632.
28. Leinonen R, Sugawara H, Shumway M, International Nucleotide Sequence Database C. The sequence read archive. Nucleic Acids Res. 2011; 39(Database issue):D19–21. doi: 10.1093/nar/gkq1019 PMID: 21062823; PubMed Central PMCID: PMC3013647.
29. Hoffmann S, Otto C, Kurtz S, Sharma CM, Khaitovich P, Vogel J, et al. Fast mapping of short sequences with mismatches, insertions and deletions using index structures. PLoS Comput Biol. 2009; 5(9):e1000502. doi: 10.1371/journal.pcbi.1000502 PMID: 19750212; PubMed Central PMCID: PMC2730575.
30. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 2014; 30(15):2114–20. doi: 10.1093/bioinformatics/btu170 PMID: 24695404; PubMed Central PMCID: PMC4103590.
31. Kingsford CL, Ayanbule K, Salzberg SL. Rapid, accurate, computational discovery of Rho-independent transcription terminators illuminates their relationship to DNA uptake. Genome Biol. 2007; 8(2):R22. doi: 10.1186/gb-2007-8-2-r22 PMID: 17313685; PubMed Central PMCID: PMC1852404.
32. Lorenz R, Bernhart SH, Honer Zu Siederdissen C, Tafer H, Flamm C, Stadler PF, et al. ViennaRNA Package 2.0. Algorithms Mol Biol. 2011; 6:26. doi: 10.1186/1748-7188-6-26 PMID: 22115189; PubMed Central PMCID: PMC3319429.
33. Garst AD, Edwards AL, Batey RT. Riboswitches: structures and mechanisms. Cold Spring Harb Perspect Biol. 2011; 3(6). doi: 10.1101/cshperspect.a003533 PMID: 20943759; PubMed Central PMCID: PMC3098680.
34. Will S, Joshi T, Hofacker IL, Stadler PF, Backofen R. LocARNA-P: accurate boundary prediction and improved detection of structural RNAs. RNA. 2012; 18(5):900–14. doi: 10.1261/rna.029041.111 PMID: 22450757; PubMed Central PMCID: PMC3334699.
35. Wolfinger MT, Svrcek-Seiler WA, Flamm C, Hofacker IL, Stadler PF. Efficient computation of RNA folding dynamics. J Phys A Math Gen. 2004; 37(17):4731–41. Pii S0305-4470(04)73887-1 doi: 10.1088/0305-4470/37/17/005 PMID: WOS:000221482800006.
36. Ikeda Y, Yagi M, Morita T, Aiba H. Hfq binding at RhlB-recognition region of RNase E is crucial for the rapid degradation of target mRNAs mediated by sRNAs in Escherichia coli. Mol Microbiol. 2011; 79(2):419–32. doi: 10.1111/j.1365-2958.2010.07454.x PMID: WOS:000286114200013.
37. Chao Y, Papenfort K, Reinhardt R, Sharma CM, Vogel J. An atlas of Hfq-bound transcripts reveals 3' UTRs as a genomic reservoir of regulatory small RNAs. EMBO J. 2012; 31(20):4005–19. doi: 10.1038/emboj.2012.229 PMID: 22922465; PubMed Central PMCID: PMC3474919.

38. Miyakoshi M, Chao Y, Vogel J. Regulatory small RNAs from the 3' regions of bacterial mRNAs. Curr Opin Microbiol. 2015; 24:132–9. doi: 10.1016/j.mib.2015.01.013 PMID: 25677420.

39. Chao Y, Vogel J. A 3' UTR-Derived Small RNA Provides the Regulatory Noncoding Arm of the Inner Membrane Stress Response. Mol Cell. 2016; 61(3):352–63. doi: 10.1016/j.molcel.2015.12.023 PMID: 26805574.

40. Miyakoshi M, Chao YJ, Vogel J. Cross talk between ABC transporter mRNAs via a target mRNA-derived sponge of the GcvB small RNA. EMBO J. 2015; 34(11):1478–92. doi: 10.15252/embj.201490546 PMID: WOS:000355996900007.

41. Kim HM, Shin JH, Cho YB, Roe JH. Inverse regulation of Fe- and Ni-containing SOD genes by a Fur family regulator Nur through small RNA processed from 3'UTR of the sodF mRNA. Nucleic Acids Res. 2014; 42(3):2003–14. doi: 10.1093/nar/gkt1071 PMID: 24234448; PubMed Central PMCID: PMC3919588.

42. Nuss AM, Heroven AK, Waldmann B, Reinkensmeier J, Jarek M, Beckstette M, et al. Transcriptomic profiling of *Yersinia pseudotuberculosis* reveals reprogramming of the Crp regulon by temperature and uncovers Crp as a master regulator of small RNAs. PLoS Genet. 2015; 11(3):e1005087. doi: 10.1371/journal.pgen.1005087 PMID: 25816203; PubMed Central PMCID: PMC4376681.

43. Raghavan R, Groisman EA, Ochman H. Genome-wide detection of novel regulatory RNAs in *E. coli*. Genome Res. 2011; 21(9):1487–97. doi: 10.1101/gr.119370.110 PMID: 21665928; PubMed Central PMCID: PMC3166833.

44. Bilusic I, Popitsch N, Rescheneder P, Schroeder R, Lybecker M. Revisiting the coding potential of the *E. coli* genome through Hfq co-immunoprecipitation. RNA Biol. 2014; 11(5):641–54. PMID: 24922322; PubMed Central PMCID: PMC4152368. doi: 10.4161/rna.29299

45. Kawano M, Reynolds AA, Miranda-Rios J, Storz G. Detection of 5'- and 3'-UTR-derived small RNAs and cis-encoded antisense RNAs in *Escherichia coli*. Nucleic Acids Res. 2005; 33(3):1040–50. doi: 10.1093/nar/gki256 PMID: 15718303; PubMed Central PMCID: PMC549416.

46. Papenfort K, Forstner KU, Cong JP, Sharma CM, Bassler BL. Differential RNA-seq of *Vibrio cholerae* identifies the VqmR small RNA as a regulator of biofilm formation. Proc Natl Acad Sci U S A. 2015; 112(7):E766–75. doi: 10.1073/pnas.1500203112 PMID: 25646441; PubMed Central PMCID: PMC4343088.

47. Loh E, Dussurget O, Gripenland J, Vaitkevicius K, Tiensuu T, Mandin P, et al. A trans-acting riboswitch controls expression of the virulence regulator PrfA in *Listeria monocytogenes*. Cell. 2009; 139(4):770–9. doi: 10.1016/j.cell.2009.08.046 PMID: 19914169.

48. Lovett PS, Rogers EJ. Ribosome regulation by the nascent peptide. Microbiol Rev. 1996; 60(2):366–85. PMID: 8801438; PubMed Central PMCID: PMC239448.

49. Henkin TM, Yanofsky C. Regulation by transcription attenuation in bacteria: how RNA provides instructions for transcription termination/antitermination decisions. Bioessays. 2002; 24(8):700–7. doi: 10.1002/bies.10125 PMID: WOS:000177092500005.

50. Remes B, Berghoff BA, Forstner KU, Klug G. Role of oxygen and the OxyR protein in the response to iron limitation in *Rhodobacter sphaeroides*. BMC Genomics. 2014; 15:794. doi: 10.1186/1471-2164-15-794 PMID: 25220182; PubMed Central PMCID: PMC4176601.

51. Glaeser J, Klug G. Photo-oxidative stress in *Rhodobacter sphaeroides*: Protective role of carotenoids and expression of selected genes. Microbiology. 2005; 151(6):1927–38.

52. Church GM, Gilbert W. Genomic sequencing. Proc Natl Acad Sci U S A. 1984; 81(7):1991–5. PMID: 6326095; PubMed Central PMCID: PMC345422.

53. Janzon L, Löfdahl S, Arvidson S. Evidence for a coordinate transcriptional control of alpha-toxin and protein a synthesis in *Staphylococcus aureus*. FEMS Microbiol Lett. 1986; 33(2–3):193–8. doi: 10.1111/j.1574-6968.1986.tb01270.x

54. Pfaffl MW. A new mathematical model for relative quantification in real-time RT-PCR. Nucleic Acids Res. 2001; 29(9):e45. PMID: 11328886; PubMed Central PMCID: PMC55695.

55. Sharma CM, Hoffmann S, Darfeuille F, Reignier J, Findeiss S, Sittka A, et al. The primary transcriptome of the major human pathogen Helicobacter pylori. Nature. 2010; 464(7286):250–5. doi: 10.1038/nature08756 PMID: 20164839.

56. Edgar R, Domrachev M, Lash AE. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Res. 2002; 30(1):207–10. PMID: 11752295; PubMed Central PMCID: PMC99122.

57. Mank NN, Berghoff BA, Hermanns YN, Klug G. Regulation of bacterial photosynthesis genes by the small noncoding RNA PcrZ. Proc Natl Acad Sci U S A. 2012; 109(40):16306–11. doi: 10.1073/pnas.1207067109 PMID: 22988125; PubMed Central PMCID: PMC3479615.

58. Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, McWilliam H, et al. Clustal W and Clustal X version 2.0. Bioinformatics. 2007; 23(21):2947–8. doi: 10.1093/bioinformatics/btm404 PMID: 17846036.

59. Sievers F, Wilm A, Dineen D, Gibson TJ, Karplus K, Li W, et al. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Mol Syst Biol. 2011; 7:539. doi: 10.1038/msb.2011.75 PMID: 21988835; PubMed Central PMCID: PMC3261699.

5.2 Processing and Decay of 6S-1 and 6S-2 RNAs in *Bacillus subtilis*

Authors: Katrin Damm, Jana Christin Wiegard, Simone Bach, Marcus Lechner, **Clemens Thölken**, Torsten Hain, Saravuth Ngo, Harald Putzer and Roland K. Hartmann

Status: manuscript in preparation

Contribution: Transcriptomic analysis, implementation of transcript length visualization

Processing and decay of 6S-1 and 6S-2 RNAs in *Bacillus subtilis*

Katrin Damm[1], Jana Christin Wiegard[2], Simone Bach[2], Marcus Lechner[2], Clemens Thölken[2], Torsten Hain[3], Saravuth Ngo[4], Harald Putzer[3] and Roland K. Hartmann[2]

[1] LOEWE Center for Synthetic Microbiology (Synmikro) & Dept. of Chemistry; Philipps-Universität Marburg; Marburg, Germany

[2] Institut für Pharmazeutische Chemie; Philipps-Universität Marburg; Marburg, Germany

[3] Institut für Medizinische Mikrobiologie; Justus-Liebig-Universität Gießen; Gießen, Germany

[4] CNRS UMR8261 (affiliated with Université de Paris Diderot, Sorbonne Paris Cité); Institut de Biologie Physico-Chimique; Paris, France

Key words: 6S RNA, non-coding RNA, pRNA, RNA degradation, endoribonuclease, RNase Y, RNase J1, RNase PH, exoribonucleases, *Bacillus subtilis*

Abstract

We analyzed processing and degradation of 6S-1 and 6S-2 RNA in *Bacillus subtilis*. Northern blot and RNA-Seq analyses using different RNase knockout strains revealed processing of precursor 6S-1 RNA (pre-6S-1 RNA) at the 5'- and 3'-end by RNases J1 and PH, respectively. 6S-1 RNA turnover proceeds primarily via the RNase Y-dependent pathway. Degradation is initiated by RNase J1-catalyzed removal of the single-stranded 5'-precursor segment to generate a monophosphorylated 5'-end, a preferred substrate for RNase Y. RNase Y then cleaves endonucleolytically in the apical loop region of 6S-1 RNA to produce 5'- and 3'-fragments of similar length that accumulate during stationary phase and which are degraded during outgrowth with participation of one or more of the four known 3'-exonucleases of *B. subtilis*. Processing of pre-6S-1 RNA is not required for product RNA (pRNA) synthesis on 6S-1 RNA as template, a mechanism which leads to the structural rearrangement of 6S-1 RNA and dissociation of 6S-1 RNA:RNA polymerase (RNAP) complexes. We infer that 5'-maturation of 6S-1 RNA via RNase J1 is not essential for the function of 6S-1 RNA, but prepares the RNA for degradation by facilitating its cleavage by RNase Y. 6S-1 RNA-derived pRNAs (~14 nt in length) are degraded by 3'- to 5'-exoribonucleases. The major 6S-2 RNA degradation pathway also involves RNase J1, Y and PH. We found that RNase Y cleaves 6S-2 RNA ~60 nt from the 5'-end and RNase J1 degrades the downstream fragment via its 5'-to-3'-exoribonuclease activity. In the absence of RNase J1, RNase Y cleaves more upstream (nt 44-50) in the 5'-central bulge, suggesting that degradosome-bound RNase J1 directs RNase Y to the region of nt 60.

Introduction

Bacterial 6S RNAs are non-coding RNAs (ncRNAs) with a size of about 200 nucleotides that interact with housekeeping RNA polymerase (RNAPs) holoenzymes to globally regulate transcription (Willkomm and Hartmann 2005; Steuten et al. 2014). Interestingly, and unlike most other bacteria, the *B. subtilis* genome harbors two 6S RNA homologs, termed 6S-1 (gene *bsrA*) and 6S-2 (gene *bsrB*) RNA (Ando et al. 2002; Suzuma et al. 2002; Trotochaud and Wassarman 2005; Beckmann et al. 2011). 6S-1 RNA, considered to be the canonical 6S RNA, reaches highest cellular levels from late exponential to stationary phase, where its levels exceed those of 6S-2 RNA at least fourfold (Hoch et al. 2015). During extended stationary phase degradation fragments of 6S-1 RNA accumulate (Beckmann et al. 2011), whose nature has been analyzed in the present study. The levels of *B. subtilis* 6S-2 RNA were reported to peak between early and mid-exponential phase (Ando et al. 2002; Barrick et al. 2005; Beckmann et al. 2011), in line with threefold higher 6S-2 versus 6S-1 RNA levels during exponential growth phase (Hoch et al. 2015). However, other studies reported the levels of 6S-2 RNA to remain quite constant throughout growth (Trotochaud and Wassarman 2005; Cavanagh et al. 2012) and to be at least twofold lower than those of 6S-1 RNA at all growth stages (Barrick et al. 2005), which may suggest that 6S-2 RNA expression is sensitive to strain background or nutrient composition. For *Escherichia coli* 6S RNA, expression levels were estimated to reach up to 10,000 molecules per cell during stationary phase (Wassarman and Storz 2000). Thus, *E. coli* 6S RNA appears to be in excess over the RNAP core enzyme (1000 - 3000 copies per cell) and σ^{70} (700 copies per cell; (Jishage and Ishihama 1995), which is in line with the experimentally determined association of up to 75% of σ^{70}-RNAP with 6S RNA during stationary phase (Wassarman and Storz 2000). The function of canonical 6S RNAs is to reprogram transcription toward stationary pase to cope with nutrient deprivation (Trotochaud and Wassarman 2004; Steuten et al. 2014) and to keep RNAP in a state where the enzyme can be instantly released from sequestration upon nutrient resupply through the synthesis of short transcripts on 6S RNA as template (see below). The function of the second 6S-2 RNA in *B. subtilis* and related bacteria is not fully clear yet. Both 6S RNAs seem to have partly overlapping functions, as a double knockout of 6S-1/2 RNAs in the *B. subtilis* PY79 background had a more severe growth phenotype than the corresponding single knockouts and the two single knockouts affected the expression of a subset of identical proteins (Hoch et al. 2015).

As a landmark feature of this ncRNA class, 6S RNAs adopt a rod-shaped structure composed of a large internal loop, termed central bulge (CB) region, which is flanked by two irregular helical stems (Barrick et al. 2005; Trotochaud and Wassarman 2005; Willkomm and Hartmann 2005). This architecture, and particularly the CB proposed to mimic an open DNA promoter, are recognized by bacterial housekeeping RNAP holoenzymes (Wassarman and Storz

2000; Trotochaud and Wassarman 2004; Gildehaus et al. 2007; Cavanagh et al. 2008). After binding to the active site of RNAP, 6S RNA itself serves as template for the transcription of short "product RNAs", called pRNAs (Wassarman and Saecker 2006). In stationary *B. subtilis* cells when nutrients including NTPs are scarce, RNAP in complex with the housekeeping sigma factor A synthesizes short pRNAs (≤ 9 nt) on 6S-1 RNA as template in an idling cycle of abortive transcription (Beckmann et al. 2011; Beckmann et al. 2012). Such abortive transcripts are thought to dissociate from 6S-1 RNA before being able to persistently rearrange the 6S-1 RNA structure (Beckmann et al. 2012). However, upon nutrient resupply when bacteria enter a new exponential growth phase, the proportion of longer pRNA transcripts (~14 nt) increases which form a stable hybrid helix with the 6S-1 RNA template RNA and persistently rearrange its structure to induce RNAP dissociation (Beckmann et al. 2011; Beckmann et al. 2012). As a result, RNAP becomes availble again for transcription initiation at chromosomal DNA promoters and the released 6S-1:pRNA complex becomes accessible to degradation (Beckmann et al. 2012).

In general, the functionality as well as the half-life of transcripts is dependent on RNA processing and decay. Surprisingly, RNA turnover differs in the well-studied model organisms *E. coli* and *B. subtilis* due to the involvement of different sets of RNases (Condon 2003; Lehnik-Habrink et al. 2012). One global player in *B. subtilis* RNA degradation is RNase Y, an endonuclease connected to the cell membrane via a membrane anchor (Lehnik-Habrink et al. 2011a; Lehnik-Habrink et al. 2011b). Acting as a functional counterpart of RNase E in *E. coli*, RNase Y initiates RNA degradation by cleavage of preferentially A/U-rich single-stranded regions (Shahbabian et al. 2009; Lehnik-Habrink et al. 2011b; Durand et al. 2012). Based on bacterial two-hybrid studies and *in vivo* crosslinking experiments it has been postulated that RNase Y binds to the exoribonucleases RNase J1 and PNPase as well as to the helicase CshA and to two glycolytic enzymes (enolase and phosphofructokinase) to form an RNA degradosome (Commichau et al. 2009; Lehnik-Habrink et al. 2010; Lehnik-Habrink et al. 2011a). Additionally, RNase J1 forms a heterotetramer with RNase J2, an inefficient 5'-exoribonuclease (Mathy et al. 2010). RNase J1, the first known bacterial 5'-to-3'-exoribonuclease (Mathy et al. 2007), was proposed to be a dual function nuclease with endo- and exonucleolytic activity modes (Li de la Sierra-Gallay et al. 2008; Newman et al. 2011). For acting as a 5'-to-3'-exoribonuclease, the enzyme prefers single-stranded 5′-extremities with monophosphorylated ends generated by a preceding endonucleolytic cleavage (Deikus et al. 2008; Yao and Bechhofer 2010) or by a pyrophosphohydrolase reaction (Richards et al. 2011). In *B. subtilis,* three pathways of RNA transcript turnover are known so far: (i) the major RNase Y-dependent pathway, (ii) 5'-directed RNase J1-dependent degradation, and (iii) an alternative RNase J1-dependent pathway involving internal cleavage by RNase J1. The major pathway is initiated by endonucleolytic RNase Y cleavage, followed by degradation of the downstream product via the 5'-to-3'-exoribonucleolytic activity of RNase J1 (Yao and Bechhofer 2010; Durand et al. 2012; Lehnik-Habrink et al. 2012).

Upstream fragments are degraded by one or more of the four known 3’-exoribonucleases (PNPase, RNase R, RNase PH and YhaM) (Craven et al. 1992; Luttinger et al. 1996; Mitra et al. 1996; Oussenko and Bechhofer 2000; Oussenko et al. 2002), of which polynucleotide phosphorylase (PNPase) is considered to be the major 3'-exonuclease (Deutscher and Reuven 1991; Oussenko et al. 2005; Liu et al. 2014). Alternatively, the pyrophosphohydrolase RppH generates a 5'-monophosphorylated RNA substrate that is degraded in 5′-to-3′-direction by RNase J1 (Richards et al. 2011; Durand et al. 2012). Only few targets are known at which J1 initiates degradation by endonucleolytic cleavage (Even et al. 2005; Deikus and Bechhofer 2009; Laalami et al. 2014). Upstream and downstream cleavage fragments are subsequently degraded in the same manner as described above for the major RNase Y-dependent pathway.

In this work we investigated the decay of *B. subtilis* 6S-1/2 RNAs and 6S-1 RNA-encoded pRNAs using a set of different strains with knockouts for RNase Y, RNases J1/J2, RppH, RNase III, RNase PH or all four known 3’-exonucleases (ΔRNase R, ΔPNPase, ΔRNase PH and ΔYhaM). We found RNases J1 and PH to be involved in 5’- and 3’-processing of 6S-1 RNA, respectively, whereas turnover of the RNA proceeds via the RNase Y-dependent pathway. We propose a model according to which 6S-1 RNA degradation is initiated by RNase J1-catalyzed 5’-to-3’-exonucleolytic or endonucleolytic removal of the single-stranded 5’-precursor segment, leaving a 5'-monophosphate end as a preferred recognition element for RNase Y. RNase Y then cleaves in the apical loop, generating 5'- and 3'-half fragments that are present at relatively high steady-state levels during stationary phase and which are further degraded by 3’- and/or 5’-exoribonucleases upon release from RNAP during a new exponential phase. In the case of 6S-2 RNA, RNase Y catalyzes the initial cleavage some distance from the 5’-end and RNase J1 degrades the downstream fragment in its 5’-3’-exoribonuclease mode.

Materials and Methods

Bacterial strains and cell culture

All knockout strains used to search for functional roles in 6S RNA processing and decay (Table 1) are derivatives of the *Bacillus subtilis* strain W168 and were provided by the laboratory of Ciaran Condon (Paris). The following concentrations of antibiotics were used for selection on agar plates or in precultures: spectinomycin 100 µg/ml; tetracycline 20 µg/ml; phleomycin 5 µg/ml; erythromycin 5 µg/ml and kanamycin 5 µg/ml. To grow strain CCB396, a quadruple mutant, a low salt LB agar (10 g/l tryptone; 5 g/l sodium chloride; 5 g/l yeast extract, adjusted to pH 7.5; 15 g/l agar) containing spectinomycin 50 µg/ml; tetracycline 10 µg/ml; phleomycin 1 µg/ml and kanamycin 5 µg/ml was used. For the control of signal specificity on Northern blots, 6S-1 and 6S-2 RNA knockout strains (*ΔbsrA, ΔbsrB*) were grown in the presence of spectinomycin 100 µg/ml or kanamycin 10 µg/ml, respectively (**Table 1**).

Cells growth was performed in LB medium at 37°C and under gentle shaking (200 rpm, Aquatron waterbath shaker, Infors AG, Germany). Complete growth curves were recorded (by optical density measurements at 600 nm, OD_{600}) using 300 ml LB medium without antibiotics after inoculation with an overnight culture (grown with antibiotics) to an OD_{600} of 0.05. To induce outgrowth, cells from extended stationary phase were diluted 1:5 in fresh prewarmed LB medium. Samples of 10 ml (for phenol extraction) or 40 ml (for TRIzol extraction) were withdrawn at different time points for RNA preparation.

RNA preparation and sequencing

Total cellular RNA was isolated from frozen cell pellets by extracting three times with hot phenol, whereas *B. subtilis* 6S-1 pRNAs (~14 nt) were extracted via the TRIzol method (Damm et al. 2015a). RNA integrity was verified on a 5% denaturing polyacrylamide (PAA) gel (8M urea) before RNA-Seq was performed. Ribosomal RNA molecules were depleted from the total RNA samples using the Ribo-Zero rRNA Removal Kit for bacteria (Epicentre). Then, the library was enriched for RNA fractions < 500 nt using the RNeasy MinElute Cleanup Kit (Qiagen) and the RNAs were subsequently treated with Tobacco Acid Pyrophosphatase (TAP, Epicentre) to convert 5'-triphosphates to 5'-monophosphates for linker ligation. Oligonucleotide adapters were ligated to the 5'- and 3'-ends of these RNA samples. First-strand cDNA synthesis was performed using Moloney Murine Leukemia Virus Reverse Transcriptase (M-MLV reverse transcriptase) and the 3'-adapter as primer. The resulting cDNAs were PCR-amplified using a high fidelity DNA polymerase. The generated cDNA was amplified by 19 PCR cycles until a DNA yield of approximately 9-26 ng/µL was reached. The PCR reaction mixtures containing the cDNA libraries were purified, using the Agencourt AMPure XP Kit (Beckmann Coulter Genomics), and pooled. Library construction was carried out at Vertis Biotechnologie AG (Freising, Germany). The paired-end sequencing reaction was conducted using an Illumina sequencer. The RNA-Seq reads were mapped to the genome of *Bacillus subtilis* subsp. subtilis str. 168 (NCBI Reference Sequence: NC_000964.3) using segemehl (Hoffmann et al. 2009) with an e-value of 0.1 and visualized via the Integrated Genome Browser (IGB)(Nicol et al. 2009) and custom scripts.

Northern blotting

For the analysis of 6S-1 and 6S-2 RNAs, 3 µg of total RNA were loaded on a 10% denaturing (8 M urea) PAA gel and Northern blotting was performed as described previously (Damm et al. 2015b). If not indicated differently, digoxigenin (DIG)-labelled antisense transcripts covering the full-length mature 6S RNA or 5S rRNA (internal loading control) were used. Antisense transcripts were synthesized by T7 RNA polymerase from linearized plasmid DNA or PCR products as templates using a DIG RNA labelling mix (Roche Diagnostics) (Damm et al. 2015b). Analysis of total RNA from *B. subtilis* strains with a 6S-1 RNA (Δ*bsrA*) or 6S-2 RNA (Δ*bsrB*) gene deletion,

respectively (Table 1, Fig. S1A), allowed us to demonstrate that the 6S-1 and 6S-2 RNA full-length probes are specific for the respective 6S RNA and fragments thereof. 6S-1 RNA precursor transcripts as well as cleavage products including the precursor were detected with a specific 5'- and 3'-labelled digoxigenin LNA/DNA probe (nt -11 to +5) (Fig. 1, Table 2). Furthermore, we designed probes against the 5'-half (nt 1-85) and 3'-half (nt 108-190) of 6S-1 RNA to assign the cleavage products to each half molecule. As markers, we prepared T7 transcripts of full-length mature (nt 1-190) 6S-1 RNA or roughly 5'- and 3'-half molecules (nt 1-85; nt 108-190) from linearized plasmid DNAs (for details, see Suppl. Material). For specific detection of 6S-1 RNA quarter parts, four LNA/DNA mixmer probes were used (Table 2; 1st quarter: nt 12-25; 2nd quarter: nt 67-80; 3rd quarter: nt 117-131; 4th quarter: nt 169-182). For pRNA detection 6 or 10 µg of total cellular RNAs were separated by 10% native PAGE (for details, see (Beckmann et al. 2010). In the case of 6S-2 RNA, we employed two probes, a full-length antisense transcript and a specific LNA/DNA mixmer against the 5'-end (nt 1-14) (Table 2).

Results

Analysis of RNases involved in 6S-1 RNA processing and decay

We analyzed 6S-1 RNA processing and decay by Northern blot experiments using probes against the full-length RNA, its 5'- and 3'-half molecules, to each quarter of the RNA and to the 5'-leader of precursor 6S-1 RNA (pre-6S-1; for details, see Fig. 1). Using total RNA isolated from stationary phase cultures of wild-type cells (wt W168) and derivative strains with RNase knockouts we observed (i) an accumulation of pre-6S-1 RNA in the RNase J1 knockout strain (Δ*rnjA*), (ii) 3'-extended pre-6S-1 RNA molecules in the RNase PH (Δ*rph*) single as well as in the related quadruple mutant with deletion of all four known *B. subtilis* 3'-exonucleases (Δ[*rph rnr yhaM pnp*]*),* and (iii) reduced amounts of 6S-1 RNA-derived fragments in the Δ*rnjA* and particularly the Δ*rny* strain (Fig. 2A; Fig. S1B). Enhanced resolution of the region comprising pre- and mature 6S-1 RNA (Fig. 2B) then revealed that 5'-maturation of 6S-1 RNA is abolished in the Δ*rnjA* strain independent of growth phase (stationary or outgrowth phase, Fig. 2B, lanes 1-3 *vs.* 4). Instead, a pre-6S-1 RNA band with its 11-nt 5'-leader shortened to roughly half its length appeared, indicating that an endoribonuclease acts on the 5'-leader with low efficiency in the absence of RNase J1. In addition, the high resolution blot confirmed two signals slightly longer than pre- and mature 6S-1 RNA, respectively, in the Δ*rph* strain relative to the parental wt strain (Fig. 2B, *cf*. lanes 4 and 5). Use of probe 6S-1_5'-half revealed that the signal area of shorter decay species (marked as "shorther fragments" in Fig. 2A and C) was partly retarded in mobility when using RNA from the Δ*rnjA* strain, demonstrating that these shifted signals represent roughly 5'-halves of 6S-1 RNA (due to endonucleolytic cleavage of 6S-1 RNA in the region of its apical loop, see

Fig. 1) that are extended by 5'-leader nucleotides (Fig. 2C, lane 2 *vs*. 1). The presence of 5'-halves with 5'-leader nucleotides in the region of shorter fragments was further confirmed by detection with the 5'-leader-specific probe (Fig. 2D, lane 2). Comparable but weaker signals obtained with RNA from strains W168 and Δ*rph* (Fig. 2D, lanes 1 and 3 *vs*. 2) suggests that some endonucleolytic cleavage of 6S-1 RNA in its apical loop region can also occur before RNase J1 has removed the 5'-leader.

We further analyzed the nature of the shorter decay intermediates using total RNA from the wt strain and probes 6S-1_5'-half and 6S-1_3'-half, unveiling that both 5'- and 3'-half fragments, about 90 nt in length (Fig. 3A, *cf.* lanes 1, 5 and 9 with the size markers in lanes 6 and 11), contribute to the signals in this area of the blot. This was confirmed using probes 6S-1_a-d (Fig. 1) specific for the individual quarters of 6S-1 RNA (Fig 3B, left four panels). Probe 6S-1_d in particular also detected even smaller fragments (Fig. 3B, lanes 17-19) that can be assigned to cleavage in the 3'-CB (~ position 140) and further downstream (around position 160). Northern blot analysis of the Δ*rny* strain using probes 6S-1_a and 6S-1_d detected faint signals in the ~ nt 90 region, suggesting that some other endonuclease can cleave in the apical loop region with low efficiency in the absence of RNase Y (Fig. S1C, lanes 2 and 7). With probe 6S-1_d, an additional signal with an estimated length of ~ 80 nt appeared in the Δ*rny* strain (Fig. S1C, lane 7), which may reflect cleavage at position 110 which is also seen in the RNA-Seq data for the Δ*rny* strain (see below, Fig. 4).

The long 6S-1 fragments (three major bands, Fig. 2A) were detectable with probes 6S-1_a-c, whereas only faint signals were obtained with probe 6S-1_d (Fig. 3B), indicating that these are primarily fragments lacking at least the 3'-terminal 20 nt of 6S-1 RNA.

RNA-Seq analysis of total RNA from wt and RNase knockout strains

For the parental W168 strain, the read coverage profile along the 6S-1 RNA sequence (Fig. 4A) and the bar diagram indicating the frequency of 5'- (green bars) and 3'-terminal read nucleotides in the libraries (Fig. 4B) clearly identified the transcription start site (at nt -11) and the mature 5'-terminus. In addition, a valley in read coverage was observed in the nt 90-110 region of libraries derived from the W168, Δ*rph* and the quadruple knockout strains (Fig. 4A) which coincided with frequent 5'-ends at nt 105-110 and frequent 3'-ends at the U-stretch of nt 88-91 (Fig. 4B). This correlates with the prominent region of shorter fragments in Northern blots, which we showed to consist of roughly 5'- and 3'-half molecules of 6S-1 RNA (Fig. 3A, middle and right panel), suggesting endonucleolytic cleavage of 6S-1 RNA in its apical loop region, possibly followed by some nibbling of the generated 5'- and/or 3'-ends. RNA-Seq further identified increased amounts of molecules with the mature 3'-end extended by 7 nt in the Δ*rph* and quadruple knockout strains (Fig. 4B), in line with the extended Northern signals in Fig. 2B (lane 5).

Surprisingly, the amount of 6S-1 RNA-specific reads was substantially reduced in the Δ*rnjA* and Δ*rny* libraries, as was the precentage of reads with 3'-ends at nt 88-91 and 5'-ends at nt 105-110 in the case of the RNase Y knockout (Fig. 4). This can be explained by RNase Y being the major endonuclease that cleaves 6S-1 RNA in the apical loop region, and RNase Y being activated by preceding RNase J1-catalyzed removal of the 5'-leader to generate a 5'-monophosphate known to be preferred by RNase Y (Shahbabian et al. 2009). For potential reasons to explain the lower number of 6S-1 RNA reads in the Δ*rnjA* and Δ*rny* libraries, see Discussion.

The Δ*rnjA*-derived library still contained a substantial proportion of 5'-ends in the nt 105-110 region (Fig. 4B). An explanation we entertain is that another nuclease than J1 or Y is able to cleave 6S-1 RNA at nt 37 and 42 in the 5'-CB (Fig. 4B). Such 5'-monophosphorylated cleavage products may then be recognized by RNase Y in the Δ*rnjA* strain, giving rise to the 5'-ends at nt 105-110. The RNA-Seq data also independently revealed the absence of 6S-1 RNA 5'-end maturation in the Δ*rnjA* strain, as well as the appearance of new RNase J1-independent endonucleolytic cleavages within the 5'-leader (*cf.* Fig. 4B, Δ*rnjA*, with Fig. 2B, lanes 1-3). The 3'-ends mapping to position 162-164 are basically found in all libraries (Fig. 4B). Combined with the finding that the long 6S-1 RNA fragments are detected by probes 6S-1_a-c, but barely by probe 6S-1_d (Fig. 3B), we conclude that a minor fraction of 6S-1 RNA molecules is initially cleaved in the 162-164 region by an yet unidentified nuclease, representing an alternative decay pathway relative to the major one involving RNase Y cleavage in the apical loop region.

Finally, the absence of substantial read numbers with 5'-ends between nt -11 and +1 in all strains except for the Δ*rnjA* mutant (Fig. 4B) either suggests that RNase J1 removes the 5'-leader in its endonucleolytic or in a rapid processive 5’-to-3’-exonucleolytic mode.

The major 5'- and 3'-ends of 6S-1 RNA-derived reads are displayed in the context of the 6S-1 RNA secondary structure in Fig. 5.

Testing for an involvement of other nucleolytic activities in 6S-1 RNA decay

We further examined a possible role of the pyrophosphohydrolase RppH as well as RNases J2 and III, the latter being a double-strand RNA-specific endonuclease (Mitra and Bechhofer 1994). The Δ*rppH* strain was included as the enzyme converts 5’-triphosphate (5'-ppp) ends to monophosphates (5'-p), which we hypothesized might favor the 5’-to-3’-exonucleolytic activity of RNase J1. Northern blot analysis of RNA from stationary phase cells of corresponding knockout strains revealed no substantial changes in the 6S-1 RNA fragment pattern relative to the parental W168 strain (Fig. S2A). Regarding RppH, this finding may indicate that removal of the 5'-triphosphate from pre-6S-1 RNA is not critical for processing by J1, which in turn might be taken as evidence for J1 removing the 5'-leader by endonucleolytic cleavage. Alternatively, another yet

unknown pyrophosphohydrolase might be the main activity that converts the 5'-triphosphate of pre-6S-1 RNA molecules to monophosphates, and J1 subsequently removes the 5'-leader exonucleolytically.

We further analyzed the RNase III deletion (Δ*rnc*) strain under outgrowth conditions, where a burst of pRNA synthesis takes place and leads to the release of 6S-1 RNA:pRNA complexes from RNAP (Beckmann et al. 2011; Beckmann et al. 2012). We considered the possibility that the released 6S-1 RNA:pRNA complex with its ~ 14 bp long hybrid helix might be a substrate for RNase III. However, the fragmentation pattern of 6S-1 RNA was unaffected in the Δ*rnc* strain under outgrowth as well as stationary phase conditions (Fig. S2B). We only observed a moderate trend toward increased intensities of 6S-1 RNA-specific signals when using total RNAs extracted from the Δ*rnc* strain (Fig. S2B, *cf.* lanes 2 and 4 with 1 and 3), but no changes in the cleavage pattern. These findings argue against a prominent role of RNase III in the degradation of 6S-1 RNA in *B. subtilis*.

The major processing and decay pathway of 6S-1 RNA

Our results obtained for 6S-1 RNA are consistent with a model according to which RNase J1 removes the 5'-leader, either via its 5'-to-3'-exonucleolytic activity or endonucleolytically, and the 3'-exonuclease RNase PH trims the 3'-end. RNase J1 processing generates mature 6S-1 RNA molecules with 5'-monophosphate ends, which then favors cleavage by RNase Y in the molecule's apical loop region (Fig. 5). It is not clear yet if the identified 5'-ends (nt 100, 105, 107-110) and 3'-ends (nt 88, 90, 91; Fig. 4B and Fig. 5) report primary cuts by RNase Y, or if the enzyme cleaves in the apical loop followed by rapid 5'- and 3'-exonucleolytic nibbling of the generated single-stranded overhangs.

6S-1 RNA - synthesis and decay of pRNAs

A key function of 6S-1 RNA is to serve as template for pRNA synthesis. We therefore asked if removal of the 5'-leader from pre-6S-1 RNA is a prerequisite for pRNA transcription. For this purpose, cells were harvested 3 min after dilution of stationary phase cells into fresh LB medium (outgrowth) to extract total RNA via the TRIzol method to enrich for small RNAs (Damm et al. 2015a). The Northern blot analysis revealed pRNA synthesis to occur with similar efficiency in the wild-type and the RNase J1 knockout strain (Fig. 6A, lanes 1 and 2). We conclude that pRNA synthesis, and by inference 6S-1 RNA function including RNAP binding, is independent of 5'-leader removal. Increased levels of 6S-1 pRNAs were detected in RNA derived from the quadruple mutant (Fig. 6B, lane 2 *vs.* 1), indicating that one or more of the 3'-to-5'-exoribonucleases are responsible for pRNA decay.

6S-2 RNA turnover in *B. subtilis*

6S-1 and 6S-2 RNA share conserved secondary structures and both function as templates for pRNA synthesis *in vitro* and *in vivo* (Burenina et al. 2014; Hoch et al. 2016), raising the question whether their decay pathways are similar. As 6S-2 RNA achieves highest levels during exponential phase while its levels drop toward stationary phase (Beckmann et al. 2011), we used total RNA from stationary phase cells to analyze 6S-2 RNA turnover. 6S-2 RNA was found to be 4- to 5-fold enriched in strains lacking RNase Y or RNase J1 (Fig. 7B, lanes 2 and 3 *vs*. 1; Fig. 8A), indicating that both enzymes are directly involved in its decay. In addition, degradation fragments were reduced in the Δ*rny* strain (Fig. 7B, shorter fragments), which was in line with the reduction of 3'-ends around nt 60 in the RNA-Seq data (Fig. 8B, *cf.* wt and Δ*rny*). In the Δ*rnjA* strain, we observed a new signal in the Northern blot (Fig. 7B, lane 3, new signal), which went along with the accumulation of 5'-ends in the Δ*rnjA* strain (Fig. 8B, positions 44-50). As a Northern probe specific for the 5'-end of 6S-2 RNA (Fig. 7C, lane 3) did not detect this new signal (Fig. 7B, lane 3), we conclude that the latter represents the 3'-product derived from cleavage in the region of nt 44-50 (Fig. 8B). Taking into account that 6S-2 RNA accumulates in the Δ*rnjA* and Δ*rny* strains, our findings suggest that 6S-2 decay is mainly initiated by endonucleolytic RNase Y cleavage in the 5'-CB (nt 44-50 or around nt 60 in the wt strain; see Discussion) followed by rapid RNase J1-catalyzed degradation of the 3'-cleavage product via its 5'-to-3'-exonucleolytic activity. Among the shorter fragments (Fig. 7B, signals a-c), only the lower band c was detected with the 5'-end specific probe (identity of band c in Fig. 7C was inferred from its migration distance relative to the full-length RNA signal, taking into account that the gels in Fig. 7B and 7C were run identically). This identifies signal c as the 5'-product derived from cleavage in the 5'-CB (~ nt 44-50), whereas signals b and c represent fragments lacking the 5'-terminal region of 6S-2 RNA. A faint signal c in RNA from Δ*rny* bacteria relative to the stronger signal of wt cells (Fig. 7C, *cf*. lanes 2 and 1) suggests that RNase Y is the major activity of the cleavages, but another endonuclease can cleave in the 5'-CB to some extent as well. The intensity reduction for signal c in the quadruple mutant (Fig. 7C, lane 5) is difficult to explain and may have indirect causes owing to the perturbation of RNA metabolism in this mutant. Finally, RNA-Seq identified 6S-2 RNAs with 3'-ends extended by three nucleotides in the *Δrph* and Δ[*rph rnr yhaM pnp*] strains (Fig. 8B). This correlates with signals "a" in Fig. 7B (lanes 4 and 5) shifted slightly upwards relative to the other strains (lanes 1-3), identifying signal "a" as a 3'-terminal fragment. We conclude that 6S-2 RNA primary transcripts of 214 nt are synthesized, which are trimmed at the 3'-end mainly by RNase PH. It should also be noted that the length of mature 6S-2 RNA in the W168 strain is 211 nt (Fig. 8) and not 203 nt as previously proposed (Ando et al. 2002).

As for 6S-1 RNA, we analyzed the effects of RppH, RNase J2 and RNase III knockouts on 6S-2 RNA degradation and did not find any evidence for a role in 6S-2 RNA decay (Fig. S2D).

Discussion

In this work, we investigated the processing and decay pathways of two non-coding RNAs, 6S-1 and 6S-2 RNA, in *B. subtilis*. In contrast to most mRNAs (Hambraeus et al. 2003), both RNAs are abundant and remarkably stable (Beckmann et al. 2011), such that cleavage products are well detectable by Northern blotting. 6S-1 RNA is transcribed as a 5'-precursor (pre-6S-1 RNA) with 11 extra nucleotides, whereas the transcription start is the mature 5'-end in the case of 6S-2 RNA. At least a substantial fraction of pre-6S-1 and 6S-2 RNAs carry 5'-ppp ends as inferred from RNA-Seq data (Beckmann et al., 2011).

Regarding 6S-1 RNA, we found RNases J1 and PH to be responsible for 5'- and 3'-processing, respectively, whereas 6S-1 RNA decay proceeds via the RNase Y-dependent pathway. 6S-1 RNA, or a fraction thereof, is evidently transcribed with a few extra nucleotides at the 3'-end, which are removed by the trimming activity of RNase PH. This might have a stabilizing effect on 6S-1 RNA, assuming that RNase PH protects the RNA from attack by other 3'-exoribonucleases that could use the 3'-overhang as starting point for more invasive degradation. According to our working model, 6S-1 RNA is processed by RNase J1 which removes the single-stranded 5'-leader to create a monophosphorylated 5'-end as a preferred substrate for RNase Y. RNase Y cleaves in the apical loop region and the resulting, roughly half-sized molecules accumulate during stationary phase ((Beckmann et al. 2011); this study). As 6S-1 RNA is thought to be protected from degradation when bound to RNAP under stationary phase conditions, we surmised that further degradation by 3'- and/or 5'-exoribonucleases or by a combination of endo- and exoribonucleases takes place in a subsequent outgrowth phase when 6S RNA dissociated from RNAP. Indeed, a corresponding accumulation of cleavage fragments during outgrowth was seen in the quadruple mutant, supporting a key role for 3'-exonucleases in this process.

One issue is the rather inefficient pre-6S-1 RNA processing by RNase J1, as pre-6S-1 RNA molecules are present at all growth stages (Beckmann et al. 2011). An explanation could be that RNase J1 removes the 5'-leader in its 5'-to-3'-exonucleolytic mode which requires the preceding conversion of the 5'-ppp to 5'-p ends. This process seems to be slow or inefficient, as substantial amounts of pre-6S-1 RNA with 5'-ppp ends are present in *B. subtilis* cells during exponential and stationary growth phases (Beckmann et al. 2011). RppH is not involved in the process (Fig. S2A), leaving the possibility that another pyrophosphohydrolase inefficiently acts on pre-6S-1 RNA (see below). We could also show that 6S-1 RNA 5'-maturation is not essential for the function of 6S-1 RNA as a template for pRNA synthesis which leads to a structural rearrangement and dissociation of the 6S-1 RNA:RNAP complex. We take this as evidence that 5'-processing by RNase J1 is

neutral to its function but plays an important role in RNA stability by initiating 6S-1 RNA decay. In accordance with this model, the levels of pre-6S-1 RNA (relative to its 5'-mature form) are lowest during extended stationary phase, likely owing to the shutdown of *de novo* 6S-1 RNA synthesis which allows RNase J1 to catch up. This then favors cleavage by RNase Y in the apical loop region, causing the 5'/3'-half fragments to accumulate (Beckmann et al. 2011), which already prepares the RNA for its final degradation by exonucleases during the next outgrowth (see above).

Remarkably, cleavage in the apical loop region is not observed *in vitro* using purified RNase Y enzyme (Fig. S3), indicating particular spatial constraints *in vivo*. The majority of 6S-1 RNA molecules is assumed to be bound to RNAP molecules in stationary phase, as inferred from rifampicin experiments that block pRNA synthesis and RNAP release as well as 6S-1 RNA decay (Beckmann et al. 2012). We thus propose that the apical loop region is the only single-stranded region that can be accessed by RNase Y in the 6S-1 RNA:RNAP complex, whereas the central bulge region of the RNA might be protected by σ^A-RNAP. Structural models predicting that the apical loop sticks out of complex with RNAP (Steuten et al. 2013) are in line with this proposal. However, there may also be other reasons why RNase Y is directed to the apical loop region *in vivo*, taking into account that we were unable to demonstrate a redirection of RNase Y cleavage to the apical loop region *in vitro* upon preincubation of 6S-1 RNA with excess amounts of σ^A-RNAP (Fig. S3).

6S-1 RNA processing by RNase J1 and decay-initiating cleavage by RNase Y may be spatially uncoupled, at least partly. RNase J1 was found to be associated with ribosomes *in vivo* (Even et al. 2005), whereas RNase Y is bound to the membrane (Hunt et al. 2006; Lehnik-Habrink et al. 2011a). It is conceivable that newly synthesized pre-6S-1 RNAs diffuse to cytoplasmic ribosome-rich areas in vicinity of the nucleoid (Bakshi et al. 2012; Mackie 2013), where 5'-processing of this fraction of pre-6S-1 RNA molecules occurs by ribosome-associated RNase J1 immediately post-transcription, long before the RNA is cut by RNase Y at the membrane-associated degradosome during stationary phase. Pre-6S-1 RNAs that capture a σ^A-RNAP holoenzyme immediately after their transcription may reach the ribosome-rich areas, and by inference RNase J1, with a time lag, which could explain the substantial steady-state levels of pre-6S-1 RNA throughout growth. In extended stationary phase, eventually all pre-6S-1 RNA molecules will diffuse to the membrane to encounter the degradosome, where RNase J1 removes the 5'-leader to enable RNase Y cleavage.

Our RNA-Seq data (Fig. 4B) and Northern blot results (Fig. 2A, Fig. S1B) revealed that cleavages in the region of the apical loop are generated primarily by RNase Y. This is consistent with previous 5'-and 3'-RACE-experiments, where the region around the apical loop was identified as a "hot spot" of endonucleolytic cleavage (Beckmann et al. 2011). However, there is another minor endonuclease activity that can also cleave in this region in the absence of RNase Y (Fig.

4B, Δ*rny*, position 110; Fig. S1C). The 5'- and 3'-ends identified (Fig. 4B) do not correspond to positions in the center of the apical loop (around position 95, Fig. 1) as expected for RNase Y which prefers single-stranded RNA regions. However, 5'- and 3'-ends between positions 94 and 95 were identified in the aforementioned RACE experiments. Thus, it is not unlikely that RNase Y initially cuts at this position (nt 94/95) to generate single-stranded overhangs that are rapidly trimmed by exoribonucleases.

Based on a previous study by Liu et al. (2014), 6S-1 RNA can be classified into the 5' = 3' category. This describes paired-end sequencing profiles for which the read levels from either end are quite equal, which is the case when upstream and downstream fragments, after endonucleolytic cleavage, are degraded by 3'-exonucleases with equal efficiency. In the case of 6S-1 RNA, adherence to the 5' = 3' profile can be attributed to the fact that the four 3'-exonucleases do not play a role in 6S-1 degradation (Fig. 4B), at least during stationary phase. In other words, in stationary phase cells, endonucleolytic cleavage fragments of 6S-1 RNA accumulate as stable intermediates (Fig. 2A).

Although processing and decay of 6S-2 RNA also involves RNases J1, Y and PH (Fig. 7D), the mechanism seems to be substantially different. In the RNA-Seq profile of the wt strain (Fig. 8A), a higher read coverage is seen in the 5'- relative to the 3'-region (5'-up category (Liu et al. 2014)). Upon cleavage in the nt 60 region, mainly by RNase Y (Fig. 8B), degradation of the 5'-cleavage fragment is evidently delayed relative to RNase J1-catalyzed degradation (see Fig. 7B) of the larger 3'-fragment. One reason for the delayed decay of the 5'-proximal fragments might be interactions with proteins that protect the RNA fragments from 3'-exonucleolytic degradation. Surprisingly, the Δ*rny* knockout, and the Δ*rnjA* knockout in particular, caused a shift to the 3'-up category, as inferred from the accumulation of 3'-fragments (Fig. 8A). This clearly illustrates that RNase J1 degrades the 3'-cleavage fragments via its 5'-to-3'-exonuclease activity. Remarkably, the endonucleolytic cleavage site (nt 55-62) is shifted to nt 44-50 in cells lacking RNase J1 (Fig. 8); cleavages in either region are substantially reduced in the Δ*rny* knockout (Fig. 8B), indicating that RNase Y acts here as the major endonuclease, but can only be inefficiently replaced with another endonuclease, as inferred from the accumulation of 6S-2 RNA reads in the Δ*rny* strain (Fig. 8A). We propose that 6S-2 RNA degradation is executed by the membrane-bound degradosome which directs RNase Y to the cleavage region of nt 55-62. However, in the absence of RNase J1, which disrupts degradosome architecture, RNase Y cleaves in the nt 44-50 region in the 5'-CB, which corresponds to the enzyme's basic substrate specificity (A/U-rich ssRNA regions; Fig S3). 6S-2 RNA decay by the membrane-bound degradosome would be in line with the fact that 6S-2 RNA is degraded toward stationary phase where the RNA is assumed to reach the membrane via diffusion, possibly after displacement from σ^A-RNAP by excess amounts of 6S-1 RNA. If correct, our model would be an example how RNase J1, when associated with RNase Y at the membrane, determines RNase Y access to an RNA substrate. This would also

be consistent with the observation that RNase Y cleavage *in vitro* not necessarily reflects the cleavage patterns seen *in vivo* (Fig. S3).

In general, a 5'-ppp terminus protects RNAs from attack by the 5'-to-3'-exoribonuclease J1 (Li de la Sierra-Gallay et al. 2008; Richards et al. 2011) and from endonucleolytic cleavage by RNase Y (Shahbabian et al. 2009). Thus, invoking some kind of pyrophosphatase in 6S-1/2 RNA degradation seems plausible. The well characterized pyrophosphohydrolase RppH of *B. subtilis*, which has no effect on 6S-1 and 6S-2 RNA decay (Fig. S2A and D), requires at least two unpaired nucleotides at the 5'-end and the second must be a G residue (Hsieh et al. 2013). A less strict preference is a purine at the third position and A favored over G at the 5'-end. Based on these constraints, RppH is predicted to be not involved in pre-6S-1 and 6S-2 RNA processing/degradation, as neither RNA carries a G at the second position. Instead, another pyrophosphohydrolase may be involved. A candidate enzyme, unidentified so far, was recently described to be relatively sequence-independent at the first three 5'-terminal positions, and estimated to be responsible for 30% of the cellular pyrophosphohydrolase activity in *B. subtilis* (Hsieh et al. 2013). Alternatively, RNase J1 may cleave the 5'-precursor of 6S-1 RNA in its endonucleolytic mode, for which the phosphorylation state has no influence on activity (Li de la Sierra-Gallay et al. 2008). According to this scenario, there might be no pyrophosphohydrolase at all involved in 6S-1/2 RNA degradation, in line with the presence of substantial cellular levels of 5'-triphosphorylated pre-6S-1 and 6S-2 RNA as inferred from RNA-Seq data (Beckmann et al. 2011).

We observed a decrease of 6S-1 and an increase of 6S-2 RNA levels in the Δ*rny* (Fig. S1B, Fig. 7B) and Δ*rnjA* strains (Fig. 2B, lane 1 *vs*. 4; Fig. 7B), which was also evident in the RNA-Seq data (Fig. 4A and 8A). This may be related to the mirror image-like expression profile of the two 6S RNAs: 6S-2 RNA accumulates in exponential phase and is degraded toward stationary phase, whereas 6S-1 RNA levels are low in early exponential phase and increase toward stationary phase. If 6S-2 RNA levels are not reduced toward stationary phase, then 6S-2 RNA may sequester a larger fraction of σ^A-RNAP enzymes in the Δ*rny* and Δ*rnjA* strains, resulting in lower transcription of genes that are activated on the way to stationary phase, which might well include the 6S-1 RNA gene. In the case of 6S-1 RNA, read profiles were 4 to 5-fold reduced for the Δ*rny* and Δ*rnjA* knockout strains (Fig. 4A), which may exceed the decrease in (pre-)6S-1 RNA levels evident from the Northern blots with RNAs derived from the two mutant strains (Fig. 2A; Fig. S1B; Fig, 2B, lane 4 vs. 1). We think that the 4 to 5-fold reduction based on the RNA-Seq read numbers correlates more closely with the Northern blot intensity reductions for 6S-1 degradation fragments in the Δ*rny* and Δ*rnjA* knockout strains. We hypothesize that fragments of 6S-1 RNA, whose intramolecular base pairing interactions can be more easily disrupted than in the context of the full-length RNA, are more efficient substrates for reverse transcription and adapter ligation than

full-length 6S-1 RNA molecules and thus enter the cDNA libraries more efficiently than intact 6S-1 RNA.

Acknowledgments

We thank Ciarán Condon for providing RNase knockout strains, and Dominik Helmecke for excellent technical assistance. This work was supported by the Deutsche Forschungsgemeinschaft (GK 1384) to K.D. and R.K.H..

References

Ando Y, Asari S, Suzuma S, Yamane K, Nakamura K. 2002. Expression of a small RNA, BS203 RNA, from the yocI-yocJ intergenic region of *Bacillus subtilis* genome. *FEMS Microbiol Lett* **207**: 29-33.

Bakshi S, Siryaporn A, Goulian M, Weisshaar JC. 2012. Superresolution imaging of ribosomes and RNA polymerase in live *Escherichia coli* cells. *Molecular microbiology* **85**: 21-38.

Barrick JE, Sudarsan N, Weinberg Z, Ruzzo WL, Breaker RR. 2005. 6S RNA is a widespread regulator of eubacterial RNA polymerase that resembles an open promoter. *RNA (New York, NY)* **11**: 774-784.

Beckmann BM, Burenina OY, Hoch PG, Kubareva EA, Sharma CM, Hartmann RK. 2011. *In vivo* and *in vitro* analysis of 6S RNA-templated short transcripts in *Bacillus subtilis*. *RNA biology* **8**: 839-849.

Beckmann BM, Grunweller A, Weber MH, Hartmann RK. 2010. Northern blot detection of endogenous small RNAs (approximately14 nt) in bacterial total RNA extracts. *Nucleic acids research* **38**: e147.

Beckmann BM, Hoch PG, Marz M, Willkomm DK, Salas M, Hartmann RK. 2012. A pRNA-induced structural rearrangement triggers 6S-1 RNA release from RNA polymerase in *Bacillus subtilis*. *The EMBO journal* **31**: 1727-1738.

Burenina OY, Hoch PG, Damm K, Salas M, Zatsepin TS, Lechner M, Oretskaya TS, Kubareva EA, Hartmann RK. 2014. Mechanistic comparison of *Bacillus subtilis* 6S-1 and 6S-2 RNAs--commonalities and differences. *RNA (New York, NY)*.

Cavanagh AT, Klocko AD, Liu X, Wassarman KM. 2008. Promoter specificity for 6S RNA regulation of transcription is determined by core promoter sequences and competition for region 4.2 of sigma70. *Molecular microbiology* **67**: 1242-1256.

Cavanagh AT, Sperger JM, Wassarman KM. 2012. Regulation of 6S RNA by pRNA synthesis is required for efficient recovery from stationary phase in *E. coli* and *B. subtilis*. *Nucleic acids research* **40**: 2234-2246.

Commichau FM, Rothe FM, Herzberg C, Wagner E, Hellwig D, Lehnik-Habrink M, Hammer E, Volker U, Stülke J. 2009. Novel activities of glycolytic enzymes in *Bacillus subtilis*: interactions with essential proteins involved in mRNA processing. *Molecular & cellular proteomics : MCP* **8**: 1350-1360.

Condon C. 2003. RNA processing and degradation in *Bacillus subtilis*. *Microbiology and molecular biology reviews : MMBR* **67**: 157-174, table of contents.

Craven MG, Henner DJ, Alessi D, Schauer AT, Ost KA, Deutscher MP, Friedman DI. 1992. Identification of the rph (RNase PH) gene of *Bacillus subtilis*: evidence for suppression of cold-sensitive mutations in *Escherichia coli*. *Journal of bacteriology* **174**: 4727-4735.

Damm K, Bach S, Müller KM, Klug G, Burenina OY, Kubareva EA, Grünweller A, Hartmann RK. 2015a. Impact of RNA isolation protocols on RNA detection by Northern blotting. *Methods in molecular biology (Clifton, NJ)* **1296**: 29-38.

-. 2015b. Improved Northern blot detection of small RNAs using EDC crosslinking and DNA/LNA probes. *Methods in molecular biology (Clifton, NJ)* **1296**: 41-51.

Deikus G, Bechhofer DH. 2009. *Bacillus subtilis* trp Leader RNA: RNase J1 endonuclease cleavage specificity and PNPase processing. *The Journal of biological chemistry* **284**: 26394-26401.

Deikus G, Condon C, Bechhofer DH. 2008. Role of *Bacillus subtilis* RNase J1 endonuclease and 5'-exonuclease activities in trp leader RNA turnover. *The Journal of biological chemistry* **283**: 17158-17167.

Deutscher MP, Reuven NB. 1991. Enzymatic basis for hydrolytic versus phosphorolytic mRNA degradation in *Escherichia coli* and *Bacillus subtilis*. *Proceedings of the National Academy of Sciences of the United States of America* **88**: 3277-3280.

Durand S, Gilet L, Bessieres P, Nicolas P, Condon C. 2012. Three essential ribonucleases-RNase Y, J1, and III-control the abundance of a majority of *Bacillus subtilis* mRNAs. *PLoS genetics* **8**: e1002520.

Even S, Pellegrini O, Zig L, Labas V, Vinh J, Brechemmier-Baey D, Putzer H. 2005. Ribonucleases J1 and J2: two novel endoribonucleases in *B.subtilis* with functional homology to *E.coli* RNase E. *Nucleic acids research* **33**: 2141-2152.

Gildehaus N, Neusser T, Wurm R, Wagner R. 2007. Studies on the function of the riboregulator 6S RNA from *E. coli*: RNA polymerase binding, inhibition of *in vitro* transcription and synthesis of RNA-directed *de novo* transcripts. *Nucleic acids research* **35**: 1885-1896.

Hambraeus G, von Wachenfeldt C, Hederstedt L. 2003. Genome-wide survey of mRNA half-lives in *Bacillus subtilis* identifies extremely stable mRNAs. *Molecular genetics and genomics : MGG* **269**: 706-714.

Hoch PG, Burenina OY, Weber MH, Elkina DA, Nesterchuk MV, Sergiev PV, Hartmann RK, Kubareva EA. 2015. Phenotypic characterization and complementation analysis of *Bacillus subtilis* 6S RNA single and double deletion mutants. *Biochimie* **117**: 87-99.

Hoch PG, Schlereth J, Lechner M, Hartmann RK. 2016. *Bacillus subtilis* 6S-2 RNA serves as a template for short transcripts *in vivo*. *RNA (New York, NY)* **22**: 614-622.

Hoffmann S, Otto C, Kurtz S, Sharma CM, Khaitovich P, Vogel J, Stadler PF, Hackermuller J. 2009. Fast mapping of short sequences with mismatches, insertions and deletions using index structures. *PLoS computational biology* **5**: e1000502.

Hsieh PK, Richards J, Liu Q, Belasco JG. 2013. Specificity of RppH-dependent RNA degradation in *Bacillus subtilis*. *Proceedings of the National Academy of Sciences of the United States of America* **110**: 8864-8869.

Hunt A, Rawlins JP, Thomaides HB, Errington J. 2006. Functional analysis of 11 putative essential genes in *Bacillus subtilis*. *Microbiology (Reading, England)* **152**: 2895-2907.

Jimenez F, Avila J, Vinuela E, Salas M. 1974. Initiation of the transcription of phi29 DNA by *Bacillus subtilis* RNA polymerase. *Biochimica et biophysica acta* **349**: 320-327.

Jishage M, Ishihama A. 1995. Regulation of RNA polymerase sigma subunit synthesis in *Escherichia coli*: intracellular levels of sigma 70 and sigma 38. *Journal of bacteriology* **177**: 6832-6835.

Laalami S, Zig L, Putzer H. 2014. Initiation of mRNA decay in bacteria. *Cellular and molecular life sciences : CMLS* **71**: 1799-1828.

Lehnik-Habrink M, Lewis RJ, Mader U, Stülke J. 2012. RNA degradation in *Bacillus subtilis*: an interplay of essential endo- and exoribonucleases. *Molecular microbiology* **84**: 1005-1017.

Lehnik-Habrink M, Newman J, Rothe FM, Solovyova AS, Rodrigues C, Herzberg C, Commichau FM, Lewis RJ, Stülke J. 2011a. RNase Y in *Bacillus subtilis*: a Natively disordered protein that is the functional equivalent of RNase E from *Escherichia coli*. *Journal of bacteriology* **193**: 5431-5441.

Lehnik-Habrink M, Pfortner H, Rempeters L, Pietack N, Herzberg C, Stülke J. 2010. The RNA degradosome in *Bacillus subtilis*: identification of CshA as the major RNA helicase in the multiprotein complex. *Molecular microbiology* **77**: 958-971.

Lehnik-Habrink M, Schaffer M, Mader U, Diethmaier C, Herzberg C, Stülke J. 2011b. RNA processing in *Bacillus subtilis*: identification of targets of the essential RNase Y. *Molecular microbiology* **81**: 1459-1473.

Li de la Sierra-Gallay I, Zig L, Jamalli A, Putzer H. 2008. Structural insights into the dual activity of RNase J. *Nature structural & molecular biology* **15**: 206-212.

Liu B, Deikus G, Bree A, Durand S, Kearns DB, Bechhofer DH. 2014. Global analysis of mRNA decay intermediates in *Bacillus subtilis* wild-type and polynucleotide phosphorylase-deletion strains. *Molecular microbiology* **94**: 41-55.

Luttinger A, Hahn J, Dubnau D. 1996. Polynucleotide phosphorylase is necessary for competence development in *Bacillus subtilis*. *Molecular microbiology* **19**: 343-356.

Mackie GA. 2013. RNase E: at the interface of bacterial RNA processing and decay. *Nature reviews Microbiology* **11**: 45-57.

Mathy N, Benard L, Pellegrini O, Daou R, Wen T, Condon C. 2007. 5'-to-3' exoribonuclease activity in bacteria: role of RNase J1 in rRNA maturation and 5' stability of mRNA. *Cell* **129**: 681-692.

Mathy N, Hebert A, Mervelet P, Benard L, Dorleans A, Li de la Sierra-Gallay I, Noirot P, Putzer H, Condon C. 2010. *Bacillus subtilis* ribonucleases J1 and J2 form a complex with altered enzyme behaviour. *Molecular microbiology* **75**: 489-498.

Mitra S, Bechhofer DH. 1994. Substrate specificity of an RNase III-like activity from *Bacillus subtilis*. *The Journal of biological chemistry* **269**: 31450-31456.

Mitra S, Hue K, Bechhofer DH. 1996. *In vitro* processing activity of *Bacillus subtilis* polynucleotide phosphorylase. *Molecular microbiology* **19**: 329-342.

Newman JA, Hewitt L, Rodrigues C, Solovyova A, Harwood CR, Lewis RJ. 2011. Unusual, dual endo- and exonuclease activity in the degradosome explained by crystal structure analysis of RNase J1. *Structure (London, England : 1993)* **19**: 1241-1251.

Nicol JW, Helt GA, Blanchard SG, Jr., Raja A, Loraine AE. 2009. The Integrated Genome Browser: free software for distribution and exploration of genome-scale datasets. *Bioinformatics (Oxford, England)* **25**: 2730-2731.

Oussenko IA, Abe T, Ujiie H, Muto A, Bechhofer DH. 2005. Participation of 3'-to-5' exoribonucleases in the turnover of *Bacillus subtilis* mRNA. *Journal of bacteriology* **187**: 2758-2767.

Oussenko IA, Bechhofer DH. 2000. The yvaJ gene of *Bacillus subtilis* encodes a 3'-to-5' exoribonuclease and is not essential in a strain lacking polynucleotide phosphorylase. *Journal of bacteriology* **182**: 2639-2642.

Oussenko IA, Sanchez R, Bechhofer DH. 2002. *Bacillus subtilis* YhaM, a member of a new family of 3'-to-5' exonucleases in gram-positive bacteria. *Journal of bacteriology* **184**: 6250-6259.

Richards J, Liu Q, Pellegrini O, Celesnik H, Yao S, Bechhofer DH, Condon C, Belasco JG. 2011. An RNA pyrophosphohydrolase triggers 5'-exonucleolytic degradation of mRNA in *Bacillus subtilis*. *Molecular cell* **43**: 940-949.

Shahbabian K, Jamalli A, Zig L, Putzer H. 2009. RNase Y, a novel endoribonuclease, initiates riboswitch turnover in *Bacillus subtilis*. *The EMBO journal* **28**: 3523-3533.

Steuten B, Hoch PG, Damm K, Schneider S, Köhler K, Wagner R, Hartmann RK. 2014. Regulation of transcription by 6S RNAs: Insights from the *Escherichia coli* and *Bacillus subtilis* model systems. *RNA biology* **11**: 508–521.

Steuten B, Setny P, Zacharias M, Wagner R. 2013. Mapping the spatial neighborhood of the regulatory 6S RNA bound to *Escherichia coli* RNA polymerase holoenzyme. *Journal of molecular biology* **425**: 3649-3661.

Suzuma S, Asari S, Bunai K, Yoshino K, Ando Y, Kakeshita H, Fujita M, Nakamura K, Yamane K. 2002. Identification and characterization of novel small RNAs in the aspS-yrvM intergenic region of the *Bacillus subtilis* genome. *Microbiology (Reading, England)* **148**: 2591-2598.

Trotochaud AE, Wassarman KM. 2004. 6S RNA function enhances long-term cell survival. *Journal of bacteriology* **186**: 4978-4985.
-. 2005. A highly conserved 6S RNA structure is required for regulation of transcription. *Nature structural & molecular biology* **12**: 313-319.
Wassarman KM, Saecker RM. 2006. Synthesis-mediated release of a small RNA inhibitor of RNA polymerase. *Science* **314**: 1601-1603.
Wassarman KM, Storz G. 2000. 6S RNA regulates *E. coli* RNA polymerase activity. *Cell* **101**: 613-623.
Willkomm DK, Hartmann RK. 2005. 6S RNA - an ancient regulator of bacterial RNA polymerase rediscovered. *Biol Chem* **386**: 1273-1277.
Yao S, Bechhofer DH. 2010. Initiation of decay of *Bacillus subtilis* rpsO mRNA by endoribonuclease RNase Y. *Journal of bacteriology* **192**: 3279-3286.

Figure legends

Fig. 1: Antisense probes used for the detection of *B. subtilis* 6S-1 RNA and fragments thereof in Northern blot experiments. The probes are illustrated as colored lines along the secondary structure of 6S-1 RNA; sky blue line: LNA/DNA mixmer probe to specifically detect 6S-1 RNA precursor molecules (probe pre-6S-1); dark blue line: antisense 6S-1 RNA (probe 6S-1_mature); orange lines: RNAs antisense to roughly the 5'-half (5', probe 6S-1_5'-half) or 3'-half (3', probe 6S-1_3'-half) of 6S-1 RNA; green lines: LNA/DNA mixmer probes (#a-d, probes 6S-1_a-d) for the detection of 6S-1 RNA fragments derived from the first to the fourth quarter of the RNA. For further details, see Materials and Methods.

Fig. 2: Northern blot analysis of 6S-1 RNA processing and decay in the *B. subtilis* wild-type (wt W168) and corresponding RNase knockout strains. (**A**) Total RNAs isolated from stationary phase cells of strains W168 (lane 1), Δ*rny* (ΔRNase Y, lane 2), Δ*rnjA* (ΔRNase J1, lane 3), Δ*rph* (RNase PH, lane 4), and the quadruple mutant (lane 5) with knockouts of *rph*, *rnr* (RNase R), *yhaM* (YhaM) and *pnp* (polynucleotide phosphorylase, PNPase), were separated by 10% denaturing PAGE; lane 6: a 197-nt T7 transcript of 6S-1 RNA loaded as size marker (M). 6S-1 RNA and 5S rRNA were probed simultaneously using full-length probes complementary to 6S-1 RNA and 5S rRNA (Fig. 1, Table 2); we confirmed in preceding Northern blot experiments using the 6S-1 or 5S RNA probe that 5S rRNA gives only rise to the single signal (indicated at the left margin) under the applied conditions. (**B**) Northern blot analysis after extended electrophoresis to resolve the region between precursor and mature 6S-1 RNA. Lanes 1-3: total RNA isolated from Δ*rnjA* bacteria in stationary (stat.) phase (lane 1) or after 3 min (lane 2) and 30 min (lane 3) of outgrowth (outgr., dilution of bacteria in fresh medium); lanes 4 and 5: total RNA isolated from stationary phase cells of strains W168 and Δ*rph*; lanes 6 and 7: T7 transcripts of mature (197 nt) and precursor (208 nt) 6S-1 RNA used as size markers (for details, see Suppl. Material). (**C**) Northern blot analysis using probe 6S-1_5'-half to determine the identity of the shorter 6S-1 RNA fragments. Lanes 1-3: as

lanes 1-3 in panel A. In addition to the 208- and 197-nt transcripts (lanes 6 and 7), we further loaded T7 transcripts mimicking the 5'- and 3'-halves of 6S-1 RNA (lanes 4 and 5; for details, see Suppl. Material). (**D**) As panel C, but using probe pre-6S-1 to specifically detect 6S-1 RNA species containing the 5'-precursor segment (11 nt; see Fig. 1). The shown blots are representative of 10 (panel A), 5 (panel B), 4 (panel C) and 7 (panel D) individual blots using 2-4 independent RNA preparations.

Fig. 3: Northern blot analysis of 6S-1 RNA processing and decay in the *B. subtilis* W168 wt strain using probes against different portions of 6S-1 RNA. (**A**) Hybridization with probes 6S-1_mature, 6S-1_5'-half and 6S-1_3'-half (see Fig. 1). (**B**) Hybridization with probes 6S-1_a to d (see Fig. 1). For T7 transcripts used as size marker and to control probe specificity, see Suppl. Material; **M** 3'-6S-1 (25 nt) and **M** 3'-6S-1 (50 nt) are DNA oligonucleotides identical in sequence to the last 25 or 50 nt of 6S-1 RNA (see Fig. 1). In the panel on the right, 12% denaturing PAGE was used to resolve fragments as short as 25 nt. The shown blots are representative of 2 individual blots performed with independent RNA preparations.

Fig. 4: RNA-Seq analysis of 6S-1 RNA processing and decay in the wt *versus* RNase mutant strains. Nucleotide coverage of cDNA reads representing 6S-1 RNA fragments visualized on top. Total RNA extracted from stationary phase cells were analyzed for the following strains: wt W168, the parental *B. subtilis* W168 strain; Δ*rph* (RNase PH knockout); Δ*rnjA* (RNase J1 knockout); Δ*rny* (RNase Y knockout) and the quadruple mutant (Δ[*rph rnr yhaM pnp*]) with knockouts of genes encoding the four 3'-exonucleases RNase PH, RNase R, YhaM and PNPase. The 6S-1 RNA sequence and corresponding nucleotide positions are indicated at the x-axis. The number of reads covering the individual nucleotide positions are indicated on the y-axis. Below, illustration of 6S-1 RNA 5'-ends or 3'-ends in percent of all 5'- and 3'-termini, respectively, found in the individual library.

Fig. 5: Illustration of the most frequent 5'- and 3'-ends of 6S-1 RNA decay intermediates (based on the RNA-Seq data shown in Fig. 4B) in the context of the 6S-1 RNA secondary structure (green nucleotides, 5'-ends; red nucleotides, 3'-ends). Also shown are the RNases identified in this study to be involved in the major processing and decay pathway of 6S-1 RNA: our results are in line with a model according to which RNase J1 removes the 5'-leader via its 5'-to-3'-exonucleolytic or endonucleolytic activity mode and the 3'-exonuclease RNase PH trims the 3'-end. RNase J1 processing generates mature 6S-1 RNA carrying a 5'-monophosphate, which favors cleavage in the apical loop region by RNase Y (gray sphere). It is not clear yet if the identified 5'-ends (nt 100, 105, 107-110) and 3'-ends (nt 88, 90, 91) report primary cuts by RNase Y, or if the enzyme

cleaves in the apical loop followed by rapid 5'- and 3'-exonucleolytic nibbling of the generated single-stranded overhangs.

Fig. 6: Northern blot analysis of 6S-1 pRNA synthesis and decay in *B. subtilis.* (**A**) Endogenous pRNAs (~14 nt), synthesized from 6S-1 RNA as template, were detected in total RNA extracts (6 µg) withdrawn from wild-type (wt W168) and Δ*rnjA* cells 3 min after induction of outgrowth. As specificity control, total RNA from a 6S-1 RNA knockout strain (*ΔbsrA*) was analyzed in parallel. A chemically synthesized pRNA 14-mer (5'-GUUCGGUCAAAACU-3') was loaded in two different amounts (0.25 and 1 ng) as length marker. (**B**) To identify RNases involved in 6S-1 pRNA decay, cells were harvested 30 min after outgrowth and 10 µg of total RNA from the wild-type (wt W168), the quadruple mutant (Δ[*rph rnr yhaM pnp*]) and the Δ*rnjA* strain were analyzed by 10% native PAGE and Northern blotting. For the *ΔbsrA* strain and markers, see legend to panel A. The shown blots are representative of 4 individual blots performed with 2-3 independent RNA preparations.

Fig. 7: Northern blot analysis of 6S-2 RNA processing and decay. (**A**) Antisense probes used in the Northern blot experiments are illustrated in the context of the putative 6S-2 RNA secondary structure. The dark blue line highlights the full-length antisense probe, and the sky blue line depicts the LNA/DNA mixmer probe 6S-2_5' (Table 2) used to detect 5'-fragments of 6S-2 RNA. (**B**) Northern blot analysis (10% denaturing PAGE) of 6S-2 RNA decay in the *B. subtilis* wild-type (wt W168) and corresponding RNase knockout strains. Total RNA was isolated from stationary phase cells of strains W168 (lane 1), Δ*rny* (ΔRNase Y, lane 2), Δ*rnjA* (ΔRNase J1, lane 3), Δ*rph* (RNase PH, lane 4), and the quadruple mutant (lane 5) with knockouts of *rph*, *rnr* (RNase R), *yhaM* (YhaM) and *pnp* (polynucleotide phosphorylase, PNPase). A 209-nt T7 transcript of 6S-2 RNA (lane 6; for details, see Suppl. Material) was loaded as size marker (**M** 6S-2 (209 nt)) and 5S rRNA was probed as loading control. Full-length antisense 6S-2 RNA (dark blue line in panel A) was used as the probe. (**C**) As panel B, but using the 5'-end-specific probe 6S-2_5' (sky blue line in panel A). (**D**) Illustration of the most frequent 5'- and 3'-ends of 6S-2 RNA decay intermediates (based on the RNA-Seq data shown in Fig. 8B); green nucleotides, 5'-ends; red nucleotides, 3'-ends. Also shown are the RNases identified in this study to be involved in 6S-2 RNA decay. The shown blots are representative of 10 (panel B) or 3 (panel C) individual blots using 2-5 independent RNA preparations.

Fig. 8: RNA-Seq analysis of 6S-2 RNA decay in the wt *versus* RNase mutant strains. Nucleotide coverage of cDNA reads representing 6S-2 RNA fragments visualised on top. Below illustration of 5'-ends and 3'-ends of 6S-2 RNA-specific reads in percent of all 5'- and 3'-termini, respectively, found in the individual library. For more details, see legend to Fig. 4.

Table 1. Strains used in this study

Strain	Genotype	Source/Reference
W168	trp +	Lab strain
CCB078	W168 *rnjB::spc*	Britton *et al.*, 2007
CCB191	W168 *rppH::spc*	Richards *et al.*, 2011
CCB322	W168 *rph::spc*	Gilet *et al.*, 2014; Oussenko *et al.*, 2005
CCB327	W168 *rnr::tc*	Oussenko *et al.*, 2005
CCB329	W168 *yhaM::Pm*	Oussenko *et al.*, 2002
CCB395	W168 *pnp::kan*	Wang *et al.*, 1996
CCB396	W168 *rph::spc rnr::tc yhaM::Pm pnp::kan*	Gilet *et al.*, 2014; Oussenko *et al.*, 2005
CCB418	W168 *txpA (-10Δ) yonT::ery rnc::spc*	Durand *et al.*,2012
CCB434	W168 *rnjA::spc*	Figaro *et al.*,2013
CCB441	W168 *rny::spc*	Figaro *et al.*,2013
MWΔ*bsrA*	PY79 *bsrA::spc*	Hoch *et al.*,2015
MWΔ*bsrB*	PY79 *bsrB::kan*	Hoch *et al.*,2015

Table 2. Northern blot probes against 6S-1, 6S-2 RNA or 5S rRNA used in this study

Probe	Sequence/Target	probe chemistry	Temperature of hybridization
6S-1_precursor	5'-DIG-ggA ctT tAt TtA aCt T-DIG-3' *	LNA/DNA mixmer	60°C
6S-1_mature	T7-transcript, nt 1-190 **	RNA	68°C
6S-1_5'-half	T7-transcript, nt 1-85 **	RNA	68°C
6S-1_3'-half	T7-transcript, nt 108-190 **	RNA	68°C
6S-1_1.quarter	5'-DIG-gGtgTacaacTaAc-DIG-3' *	LNA/DNA mixmer	60°C
6S-1_2.quarter	5'-DIG-gtaCgcCaTttAaa-DIG-3' *	LNA/DNA mixmer	60°C
6S-1_3.quarter	5'-DIG-gtGcccTctTttAaa-DIG-3' *	LNA/DNA mixmer	60°C
6S-1_4.quarter	5'-DIG-aAtAgTgccgTtgc-DIG-3' *	LNA/DNA mixmer	60°C
6S-1_pRNA	5'-DIG-aGttTtgAccGaAc-3' *	LNA/DNA mixmer	50°C
6S-2_mature	T7-transcript, nt 1-203 **	RNA	68°C
6S-2_5'	5'-DIG-cacAaAgtAgctTc-3' *	LNA/DNA mixmer	55°C
5S_mature	T7-transcript, nt 1-115 **	RNA	68°C

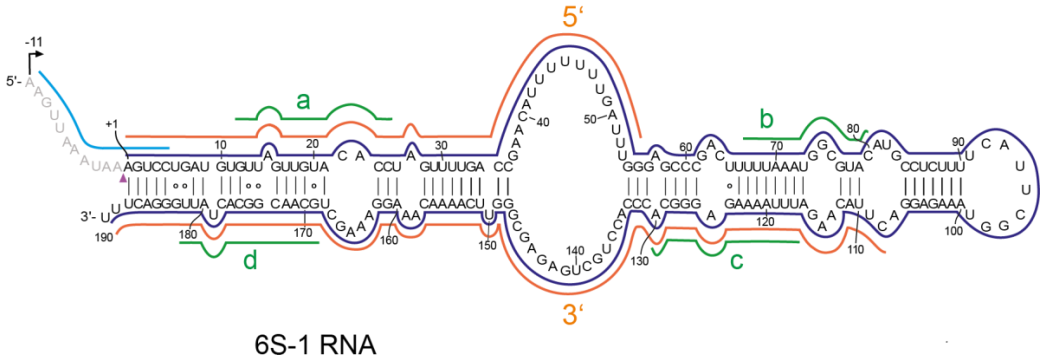
* upper case letters depict LNA and lower case letters DNA, Exiqon

** antisense transcripts covering the region as stated, internally labelled with digoxigenin-UTP

Table 3. RNA-Seq- Read numbers of 6S-1 and 6S-2 RNA

Strain	Reads	6S-1	6S-2	6S-1 %	6S-2 %	6S-1/6S-2
wt *W168*	13292751	795689	33011	5.99	0.25	24.1
Δrph	15238650	875161	33408	5.74	0.22	26.2
ΔrnjA	15128998	175442	162755	1.16	1.08	1.1
Δrny	13390664	172630	155920	1.29	1.16	1.1
Δ*(rph rnr yhaM pnp)*	12814465	600173	13818	4.68	0.11	43.4

Figures



6S-1 RNA

Fig.1

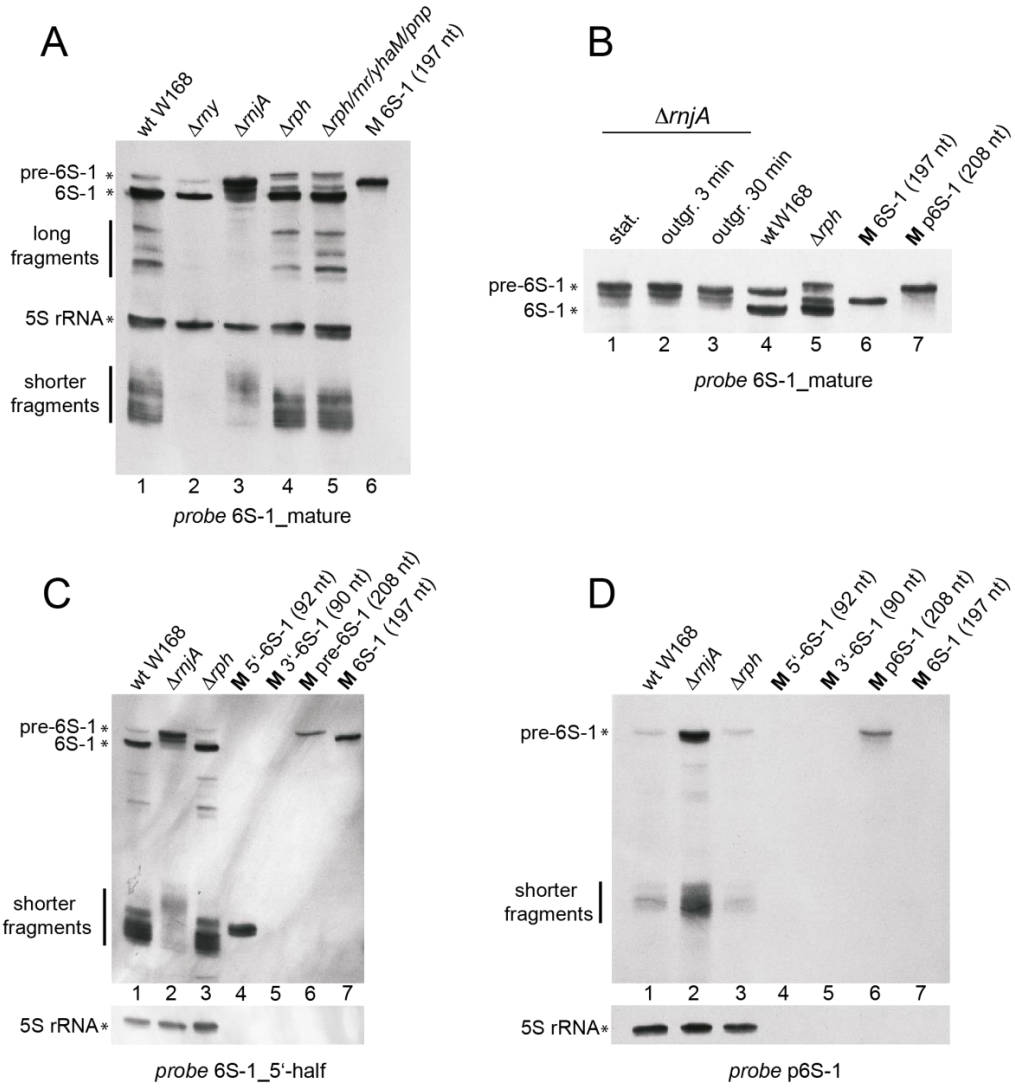

A
wt W168
Δrny
ΔrnjA
Δrph
Δrph/rnr/yhaM/pnp
M 6S-1 (197 nt)
pre-6S-1 *
6S-1 *
long fragments
5S rRNA *
shorter fragments
1 2 3 4 5 6
probe 6S-1_mature
B
ΔrnjA
stat.
outgr. 3 min
outgr. 30 min
wtW168
Δrph
M 6S-1 (197 nt)
M p6S-1 (208 nt)
pre-6S-1 *
6S-1 *
1 2 3 4 5 6 7
probe 6S-1_mature
C
wt W168
ΔrnjA
Δrph
M 5'-6S-1 (92 nt)
M 3'-6S-1 (90 nt)
M pre-6S-1 (208 nt)
M 6S-1 (197 nt)
pre-6S-1 *
6S-1 *
shorter fragments
1 2 3 4 5 6 7
5S rRNA *
probe 6S-1_5'-half
D
wt W168
ΔrnjA
Δrph
M 5'-6S-1 (92 nt)
M 3'-6S-1 (90 nt)
M p6S-1 (208 nt)
M 6S-1 (197 nt)
pre-6S-1 *
shorter fragments
1 2 3 4 5 6 7
5S rRNA *
probe p6S-1

Fig. 2

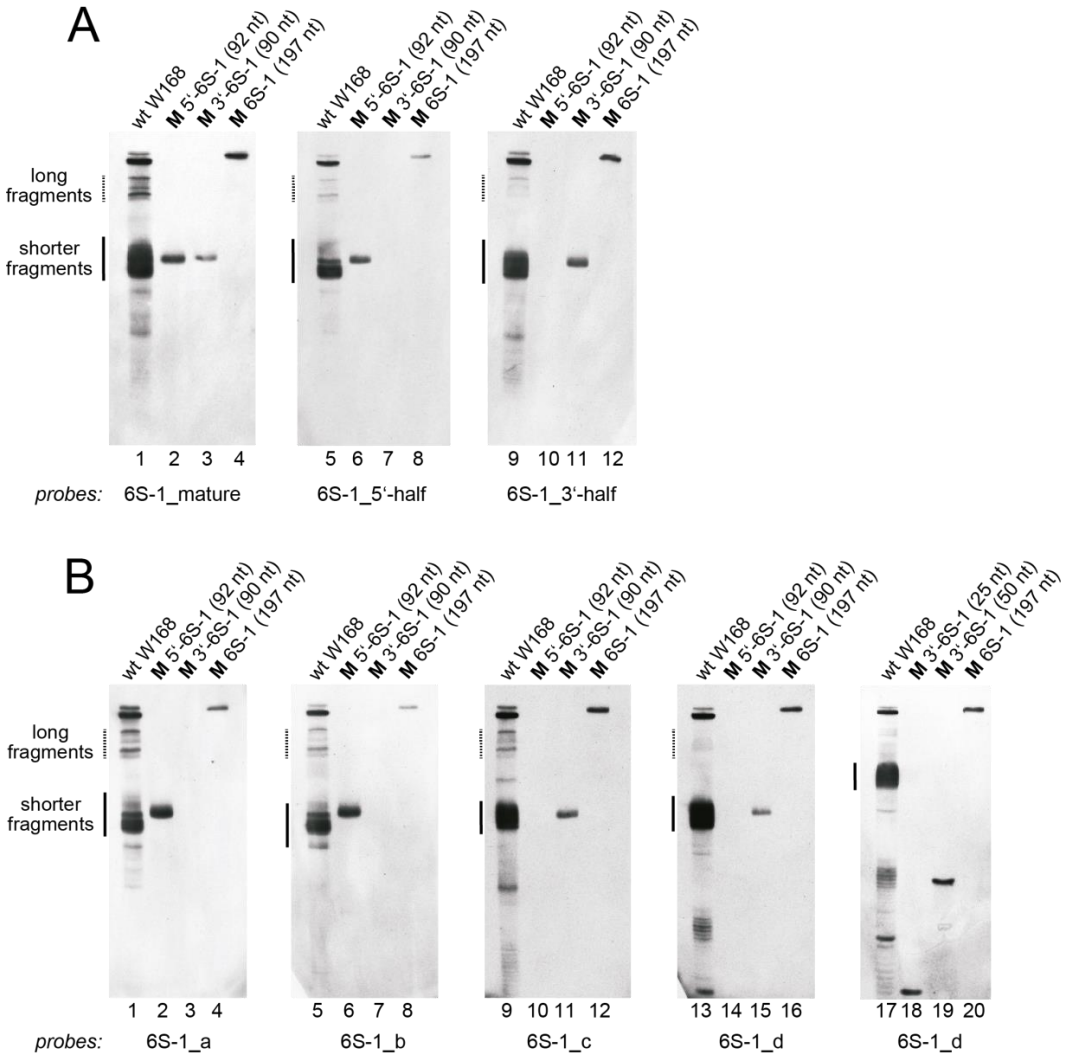

A
wt W168
M 5'-6S-1 (92 nt)
M 3'-6S-1 (90 nt)
M 6S-1 (197 nt)
long fragments
shorter fragments
1 2 3 4
5 6 7 8
9 10 11 12
probes:
6S-1_mature
6S-1_5'-half
6S-1_3'-half
B
wt W168
M 5'-6S-1 (92 nt)
M 3'-6S-1 (90 nt)
M 6S-1 (197 nt)
M 3'-6S-1 (25 nt)
M 3'-6S-1 (50 nt)
long fragments
shorter fragments
1 2 3 4
5 6 7 8
9 10 11 12
13 14 15 16
17 18 19 20
probes:
6S-1_a
6S-1_b
6S-1_c
6S-1_d
6S-1_d

Fig. 3

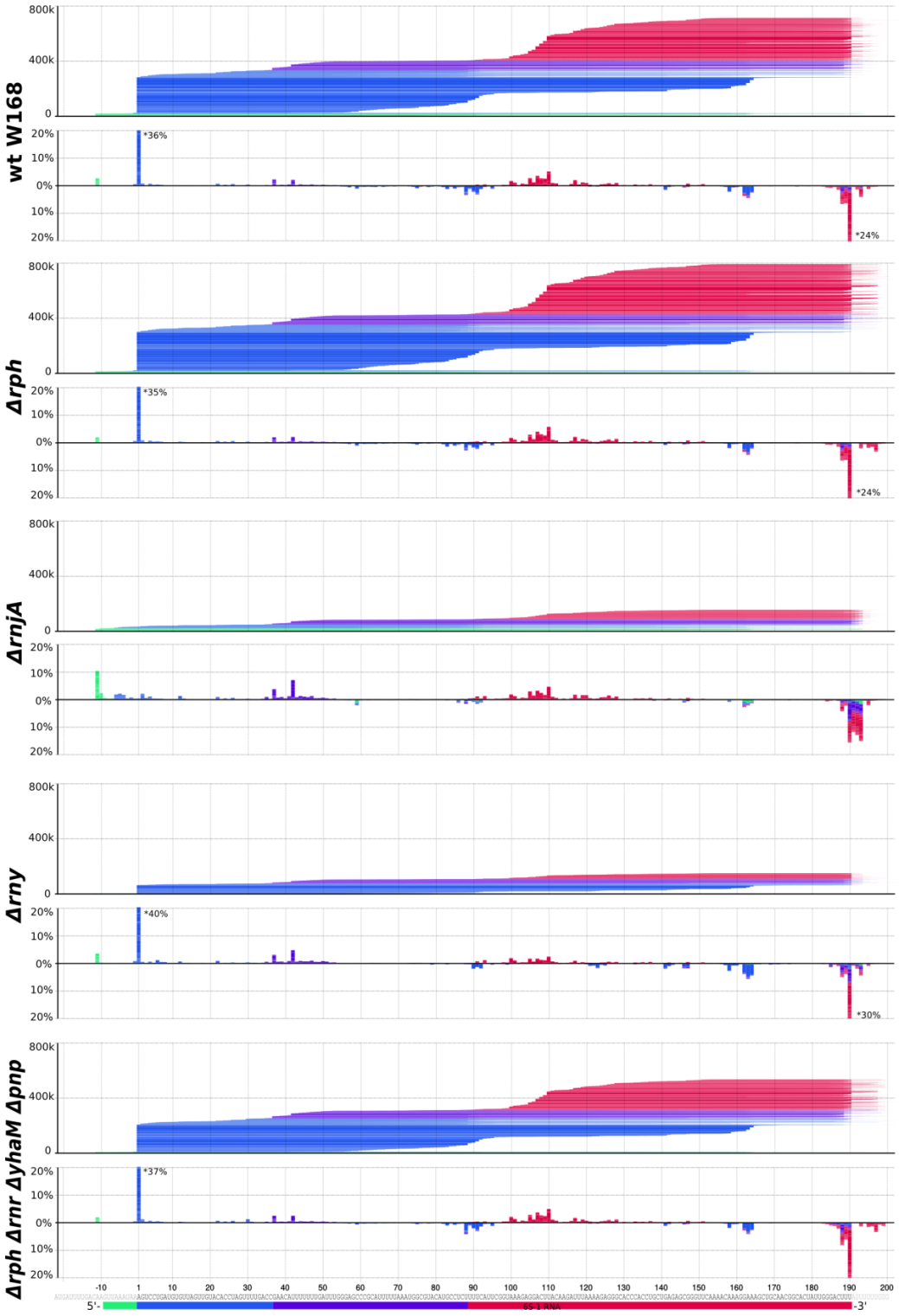

wt W168
Δrph
ΔrnjA
Δrny
Δrph Δrnr ΔyhaM Δpnp
800k
400k
0
20%
10%
0%
*36%
*24%
*35%
*40%
*30%
*37%
-10
1
10
20
30
40
50
60
70
80
90
100
110
120
130
140
150
160
170
180
190
200
5'-
6S-1 RNA
-3'

Fig. 4

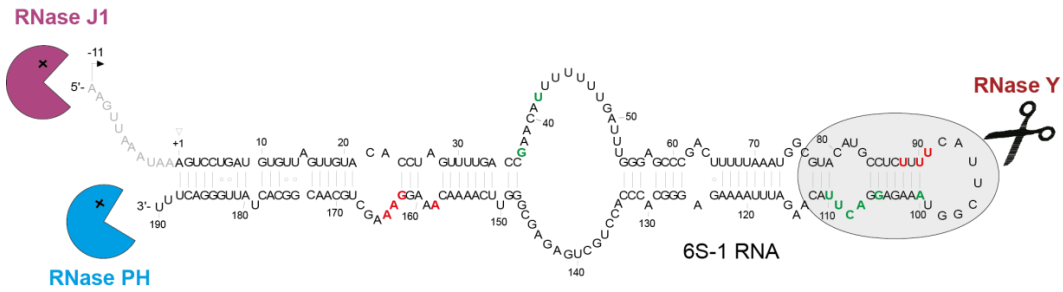

RNase J1
RNase Y
RNase PH
6S-1 RNA

Fig. 5

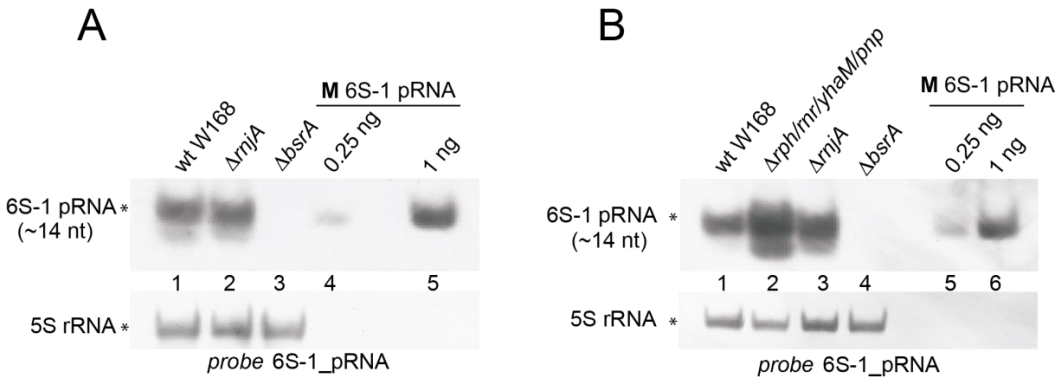

A
M 6S-1 pRNA
wt W168
ΔrnjA
ΔbsrA
0.25 ng
1 ng
6S-1 pRNA *
(~14 nt)
5S rRNA *
1 2 3 4 5
probe 6S-1_pRNA
B
M 6S-1 pRNA
wt W168
Δrph/rnr/yhaM/pnp
ΔrnjA
ΔbsrA
0.25 ng
1 ng
6S-1 pRNA *
(~14 nt)
5S rRNA *
1 2 3 4 5 6
probe 6S-1_pRNA

Fig. 6

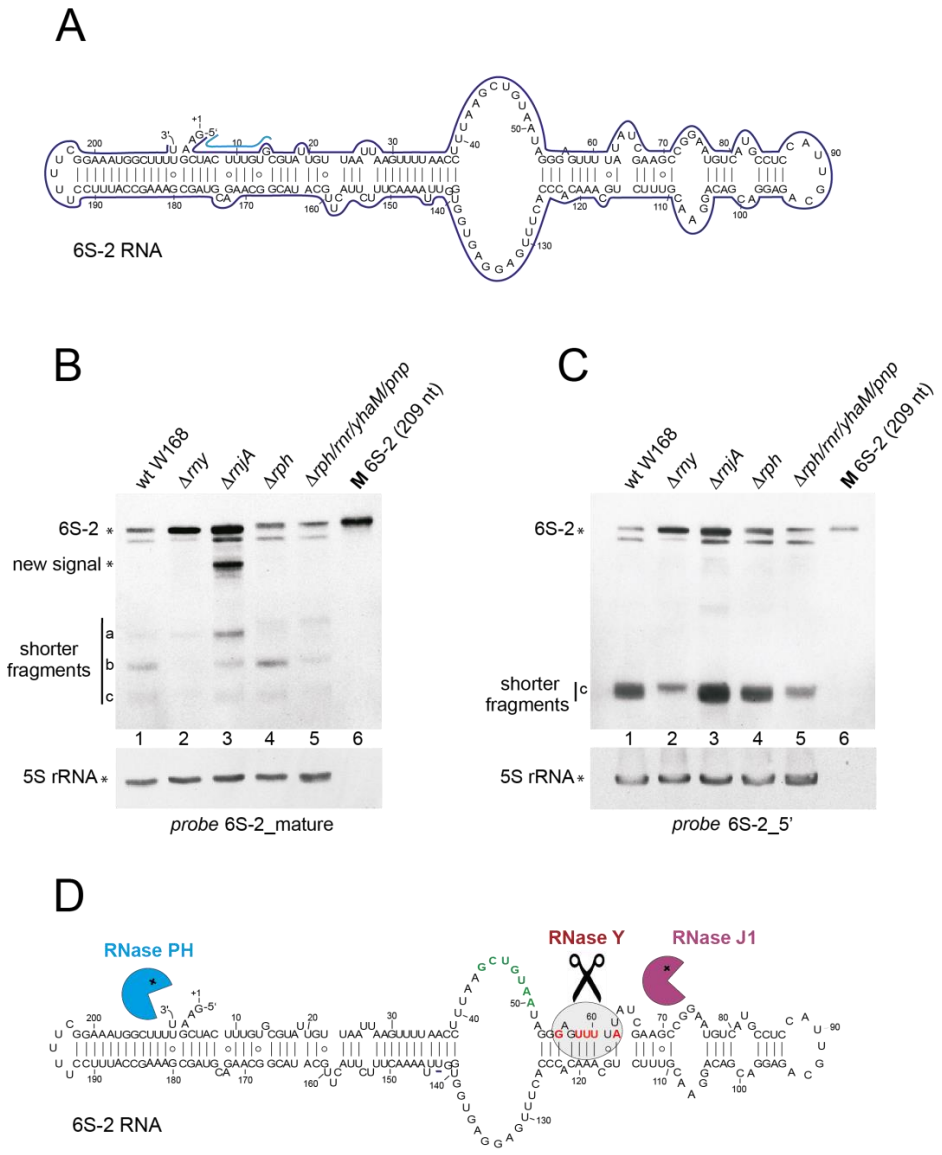

A
6S-2 RNA
B
wt W168
Δrny
ΔrnjA
Δrph
Δrph/rnr/yhaM/pnp
M 6S-2 (209 nt)
6S-2 *
new signal *
shorter fragments
a
b
c
1 2 3 4 5 6
5S rRNA *
probe 6S-2_mature
C
wt W168
Δrny
ΔrnjA
Δrph
Δrph/rnr/yhaM/pnp
M 6S-2 (209 nt)
6S-2 *
shorter fragments
c
1 2 3 4 5 6
5S rRNA *
probe 6S-2_5'
D
RNase PH
RNase Y
RNase J1
6S-2 RNA

Fig. 7

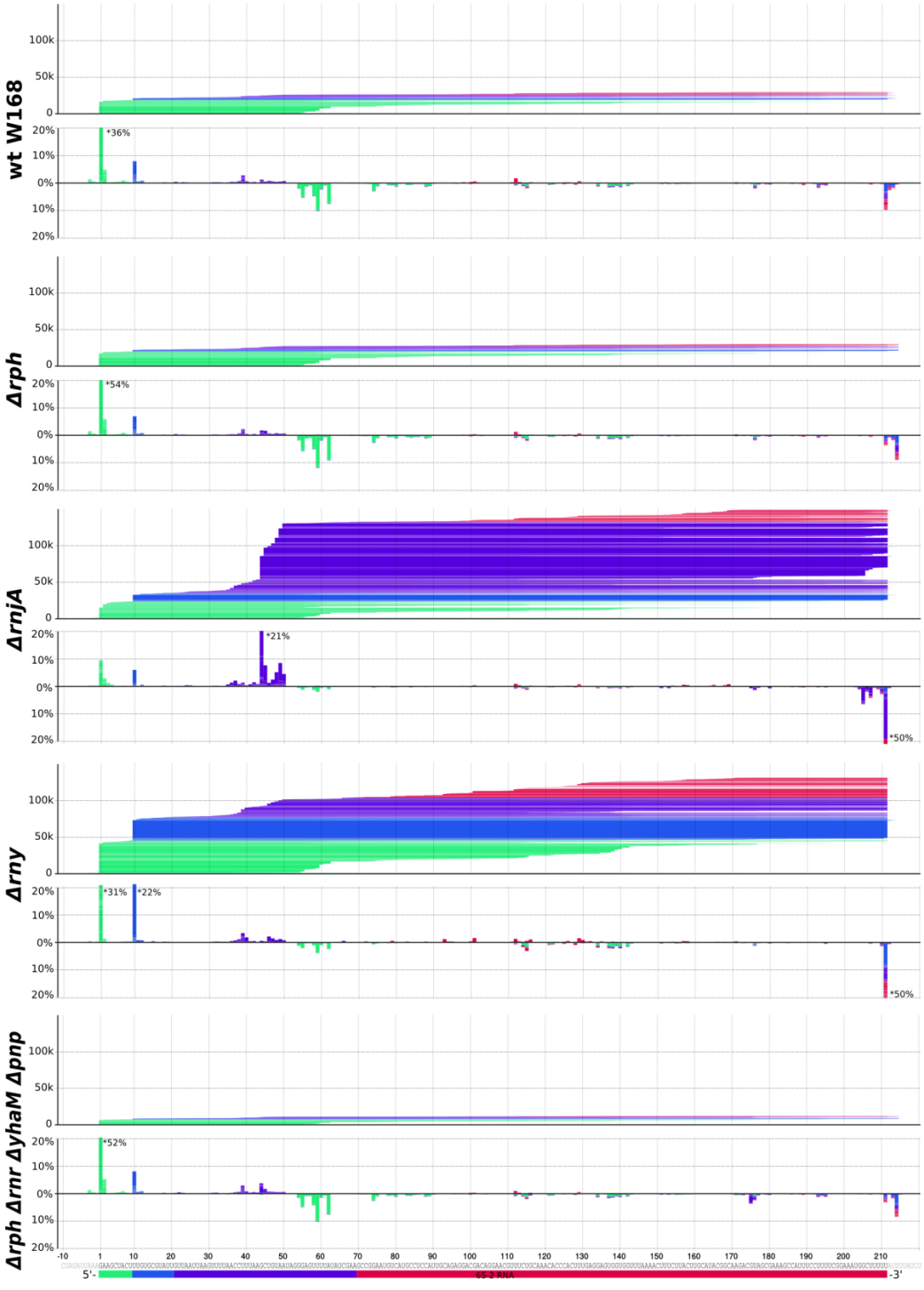

wt W168
Δrph
ΔrnjA
Δrny
Δrph Δrnr ΔyhaM Δpnp
100k
50k
0
20%
10%
0%
*36%
*54%
*21%
*50%
*31%
*22%
*52%
-10
1
10
20
30
40
50
60
70
80
90
100
110
120
130
140
150
160
170
180
190
200
210
5'-
6S-2 RNA
-3'

Fig. 8

5.3 Distribution of Ribonucleoprotein and Protein-only RNase P in Eukarya

Authors: Marcus Lechner, Walter Rossmanith, Roland K. Hartmann, **Clemens Thölken**, Bernard Gutmann, Philippe Giegé and Anthony Gobert

Journal: Molecular biology and evolution

DOI: 10.1093/molbev/msv187

Contributions: Structural alignment of RNase P RNAs, phylogenetic tree of PRORP

Distribution of Ribonucleoprotein and Protein-Only RNase P in Eukarya

Marcus Lechner,[1] Walter Rossmanith,[2] Roland K. Hartmann,[1] Clemens Thölken,[1] Bernard Gutmann,[3] Philippe Giegé,*[,3] and Anthony Gobert*[,3]

[1]Institut für Pharmazeutische Chemie, Philipps-Universität Marburg, Marburg, Germany

[2]Zentrum für Anatomie & Zellbiologie, Medizinische Universität Wien, Wien, Austria

[3]Institut de Biologie Moléculaire des Plantes du CNRS, Strasbourg, France

***Corresponding author:** E-mail: anthony.gobert@ibmp-cnrs.unistra.fr; giege@unistra.fr.

Associate editor: Claus Wilke

Abstract

RNase P is the endonuclease that removes 5′ leader sequences from tRNA precursors. In Eukarya, separate RNase P activities exist in the nucleus and mitochondria/plastids. Although all RNase P enzymes catalyze the same reaction, the different architectures found in Eukarya range from ribonucleoprotein (RNP) enzymes with a catalytic RNA and up to 10 protein subunits to single-subunit protein-only RNase P (PRORP) enzymes. Here, analysis of the phylogenetic distribution of RNP and PRORP enzymes in Eukarya revealed 1) a wealth of novel P RNAs in previously unexplored phylogenetic branches and 2) that PRORP enzymes are more widespread than previously appreciated, found in four of the five eukaryal supergroups, in the nuclei and/or organelles. Intriguingly, the occurrence of RNP RNase P and PRORP seems mutually exclusive in genetic compartments of modern Eukarya. Our comparative analysis provides a global picture of the evolution and diversification of RNase P throughout Eukarya.

Key words: **RNase P, eukaryal evolution, PRORP, ribonucleoproteins, tRNA biogenesis.**

Letter

RNase P is the endonuclease that removes 5′ leader sequences from tRNA precursors, an essential step in tRNA maturation (Lai et al. 2010; Liu and Altman 2010). The virtually ubiquitous enzyme independently originated at least twice in evolution with different architectures. Ribonucleoprotein (RNP) enzymes based on a catalytic RNA molecule (P RNA) represent the more ancient type that is found in all three domains of life. Although their RNA is structurally conserved, their protein partners are highly divergent with a single protein in Bacteria, 4–5 in Archaea, and up to 10 in eukaryal nuclei (Hartmann et al. 2009; Ellis and Brown 2010; Lai et al. 2010; Liu and Altman 2010; Walker et al. 2010). All known nuclear RNase P RNPs are composed of a P RNA of about 350 nt and a set of proteins, always including RPP21/RPR2, RPP29/POP4, RPP30/RPP1, POP5, POP1, RPP20/POP7, and RPP25/POP6 (Hartmann E and Hartmann RK 2003; Rosenblad et al. 2006; Walker et al. 2010). The reasons for the massive increase in the protein moiety of the enzyme in Eukarya as compared with Archaea or Bacteria are poorly understood and have been speculated to be related to added functionality of the eukaryal enzyme (Marvin and Engelke 2009a, 2009b; Jarrous and Gopalan 2010), although recent RNase P replacement experiments do not support such notion (Weber et al. 2014). Studies of the prevalence of nuclear RNP RNase P subunits in eukaryal genomes are complicated by the presence of a related RNP, RNase MRP, exclusively found in Eukarya and involved in 5.8S rRNA maturation. This RNP enzyme is composed of a structurally related, but nonetheless distinguishable RNA, and a largely overlapping set of proteins (Jarrous and Gopalan 2010; Walker et al. 2010). In fact, it appears that RPP21 is the only protein not shared by the two RNPs, but consistently specific to RNase P.

A fundamentally different type of RNase P is composed of protein only (PROteinaceous RNase P, PRORP) and appears confined to the eukaryal domain. In its simplest form it consists of a single 60-kDa protein, but requires additional subunits in some cases, for example, two other protein components in human mitochondrial RNase P (Holzmann et al. 2008; Gobert et al. 2010, 2013; Gutmann et al. 2012; Taschner et al. 2012; Pinker et al. 2013). The two kinds of RNase P are highly similar in terms of substrate and cleavage specificity, and they were even found to be functionally exchangeable in *Escherichia coli* and *Saccharomyces cerevisiae* (Gobert et al. 2010; Taschner et al. 2012; Weber et al. 2014).

The discovery of protein-only RNase P (PRORP) enzymes in Eukarya pointed out that the evolution of RNase P is more intriguing and complex than previously thought. Questions are raised as to when PRORP appeared during evolution, and if there may still be evolutionary traces of its coexistence with RNP RNase P within the same cellular compartment. Where and how did such a coexistence lead to the divergent specialization and compartmentalization of the different RNase P enzymes? Here, we analyze and compare the prevalence and architectural type of both RNP and PRORP enzymes in Eukarya. We find that PRORP enzymes are widespread among eukaryal lineages and propose reasonable scenarios for the evolution of RNase P in Eukarya.

Results and Discussion

Incidence of Nuclear Ribonucleoprotein RNase P

Here we update the distribution of P RNA and RPP21 (the protein subunit not found in RNase MRP) in eukaryal nuclear genomes, based on previously published studies (Hartmann E and Hartmann RK 2003; Marquez et al. 2005; Piccinelli et al. 2005; Rosenblad et al. 2006) and analyses of newly available genome data, to determine the prevalence of nuclear RNP RNase P in the different branches of Eukarya. The results are summarized in table 1 and the inventory detailed in supplementary table S1, Supplementary Material online (http://bioinf.pharmazie.uni-marburg.de/supplements/rnase_p_2015/ last accessed September 14, 2015). For example, we identified a variety of novel P RNAs including hitherto unexplored taxa.

In brief, a P RNA and RPP21 are prevalent among the Holozoa subgroup of Opisthokonta. Within metazoans, P RNA candidates were newly identified in the more basal Placozoa, Porifera, and in radially symmetric animals (supplementary figs. S1–S5, Supplementary Material online). In Nucletmycea, P RNAs are identifiable in all branches except for Nucleria. Among Amoebozoa, nuclear RNP RNase P is generally present. Relative to previous analyses (Marquez et al. 2005; Piccinelli et al. 2005), we predicted additional P RNAs and RPP21 homologs in Archamoebae and Dictyostelia. In contrast, within the photosynthetic supergroup Archaeplastida (plants and algae with chloroplasts of primary endosymbiotic origin), RNP RNase P appears absent from the nuclei of Chloroplastida. However, P RNAs are predicted in glaucophytes and in rhodophytes. In the SAR (Stramenopiles, Alveolata, Rhizaria) group, P RNA and RPP21 were not identified in Stramenopiles, consistent with previous studies (Hartmann E and Hartmann RK 2003; Piccinelli et al. 2005; Rosenblad et al. 2006), but were found in Ciliophora and Apicomplexa genomes (Alveolata). In Excavata, the occurrence of nuclear RNP RNase P is widespread, but appears to have been lost in Euglenozoa. In Haptophyta and Cryptophyceae, P RNA or RPP21 could not be identified; yet, genome information is scarce in these clades and it remains unclear whether this is due to the loss of RNP RNase P or to structurally highly deviant P RNA and RPP21 homologs.

Incidence of Organellar Ribonucleoprotein RNase P

Mitochondria (mt) and plastids (pls) possess their own genome coding for a complete or partial set of tRNAs. They originated from primary endosymbiosis with an ancestral α-proteobacterium and a cyanobacterium, respectively, yet pls also derive from secondary or tertiary endosymbiosis in various groups. It is thus not surprising to find bacterial-like P RNAs still encoded in some organellar genomes. Organelle RNP RNase P, however, is particularly diverse (Rossmanith 2012) and P RNAs are highly degenerate in some cases (Seif et al. 2005). We have (re)analyzed the occurrence of mt and pl–P RNAs in organelle genomes throughout Eukarya as well as the occurrence of RnpA and Rpm2, two proteins of organellar RNP RNase P. The comprehensive list of all identified organellar P RNAs and proteins is given in supplementary table S1, Supplementary Material online, and summarized in table 1.

In short, in the supergroup Opisthokonta, no P RNA gene was found in the mitochondrial genomes of Holozoa. Most mitochondrial genes were found in the fungal lineage particularly in saccharomycetaceaen species. Among Archaeplastida, a patchy occurrence of P RNAs was found in organellar genomes of phylogenetically basal alga including Glaucophyta, Rhodophyceae, and Chlorophyta. No P RNA gene was found in Streptophyta. Most, if not all, pl-encoded P RNAs were found in primary photosynthetic Eukarya. In Excavata, P RNAs were only found in jakobid mtDNAs (Seif et al. 2006). Finally, in the groups of amoebozoa and SAR, organellar P RNA appears to be scarce. All in all, organelle P RNA occurrence is patchy. In some phyla, they were either lost or their sequences have diverged to an extent that makes them undetectable by recognition algorithms used here. Protein subunits of these enzymes are even more elusive. The subunits previously identified are bacterial-type RNase P proteins (RnpA) and a pentatricopeptide repeat (PPR) protein called Rpm2, both nuclear encoded and unrelated to PRORP. Within the fungal branch, Rpm2 was shown to be part of mitochondrial RNase P in *S. cerevisiae* (Morales et al. 1992; Daoud et al. 2012). Close Rpm2 homologs are only found in Saccharomycetales (supplementary fig. S6, Supplementary Material online). In Archaeplastida, no P protein of bacterial origin is encoded in any organellar genome, although *rnpA*-like genes are encoded in several nuclear genomes in Mammiellophyceae of the Chlorophyta subgroup, (Lai et al. 2011) and these RnpA proteins are predicted to localize to organelles (supplementary table S2, Supplementary Material online). Our analysis and three-dimensional structure predictions revealed that these algae RnpAs are characterized by N- and C-terminal extensions not present in bacterial RnpAs (supplementary figs. S7 and S8, Supplementary Material online). Their function is unknown, but might be involved in specific contacts with algae organellar P RNAs or with yet unidentified proteins.

Incidence of Protein-Only RNase P

Our analyses confirm and substantiate previous observations that a number of eukaryal groups lack RNase P genes for a nuclear and/or organellar RNP enzyme. We thus performed a systematic analysis of the distribution and localization of putative PRORP enzymes to determine whether PRORP could be the RNase P in these lineages/compartments. As a prerequisite, we had to define robust features characterizing PRORP. Candidates were only considered as genuine PRORPs when their architecture included a specific C-terminal NYN (N4BP1, YacP-like Nuclease) metallonuclease domain presumably originating from the bacterial ribonuclease *yacP* (Anantharaman and Aravind 2006), an N-terminal α-super helical domain containing PPR motifs (Small et al. 2004) as revealed by systematic structure predictions and a bipartite zinc-binding module connecting the two main domains. Further signatures are present in specific phyla. Their occurrence might point out additionally acquired functions or

Table 1. Overview of the occurrence of RNP and PRORP RNase P enzymes in nuclei and organelles throughout the eukaryal tree.

						Nucleus Encoded				Organelle Encoded		
						Nuclear RNase P				Organellar RNase P		
						RNP RNase P		PRORP		RNP RNase P		
Supergroups	Subgroups				Representative species	P RNA	P protein	Nuclear	Organellar	P protein	P RNA	
Opisthokonta	Holozoa	Metazoa	Animalia	Bilateria	*Homo sapiens*	n[a]	n[a]		m[a]			C
				Radiata	*Acropora digitifera*	n			m			C
			Porifera		*Amphimedon queenslandica*	n	n		?			
			Placozoa		*Trichoplax adhaerens*	n	n		m			C
		Choanomonada			*Monosiga brevicollis*	n			m			C
		Filasterea			*Capsaspora owczarzaki*	n			m			C
		Ichthyosporea			*Amoebidium parasiticum*	n	n		m			C
					Sphaeroforma arctica			n	m			C
	Nucletmycea	Nuclearia			*Nuclearia simplex*							
		Fungi	Microsporidia		*Encephalitozoon cuniculi*	n	n					C
			Chytridiomycota		*Spizellomyces punctatus*	n–	n					
			Blastocladiales		*Allomyces macrogynus*	n						
			Mucoromycotina		*Rhizopus oryzae*	n	n				m	C
			Mortierellaceae		*Mortierella verticillata*	n	n				m	C
			Dikarya	Ascomycota	*Saccharomyces cerevisiae*	n[a]	n[a]			(m)[a]	(m)[a]	C
				Basidiomycota	*Postia placenta*	n	n				(m)[a]	C
Amoebozoa	Discosea				*Acanthamoeba castellanii*	n	n					
	Archamoebae				*Entamoeba dispar*	n						C
	Myxogastria				*Physarum polycephalum*	n						
	Dictyostelia				*Dictyostelium discoideum*	n[a]	n[a]					
Archaeplastida	Glaucophyta				*Cyanophora paradoxa*	n′	n				p	
	Rhodophyceae	Bangiales			*Porphyra purpurea*						p	
		Cyanidiales			*Cyanidioschyzon merolae*						p	
		Florideophycidae			*Chondrus crispus*						p	
		Porphyridiophyceae			*Porphyridium purpureum*	n′					p	
	Chloroplastida (Viridiplantae)	Chlorophyta	Trebouxiophyceae		*Chlorella variabilis*			n	m, p			C
			Chlorophyceae		*Chlamydomonas reinhardtii*			n	m, p			C
			Mamiellophyceae		*Ostreococcus tauri*			n[a]		m, p	m, p	C
		Charophyta	Streptophyta		*Arabidopsis thaliana*			n[a]	m, p[a]			C
SAR	Stramenopiles	Blastocystis			*Blastocystis hominis*				?			
		Labyrinthulomycetes			*Schizochytrium aggregatum*				m			
		Pelagophyceae			*Aureococcus anophagefferens*			n	m, p			C
		Eustigmatales			*Nannochloropsis gaditana*			n	m, p			C
		Peronosporomycetes			*Phytophthora sojae*			n	m			C
		Phaeophyceae			*Ectocarpus siliculosus*			n	m, p			C
		Diatomea			*Thalassiosira pseudonana*			n	m, p			C
	Alveolata	Protalveolata			*Perkinsus marinus*				?			
		Dinoflagellata			*Karenia brevis*				?			
		Apicomplexa		Haemosporidia	*Plasmodium falciparum*	n	n		a			C
		Ciliophora			*Tetrahymena thermophila*	n						
	Rhizaria	Cercozoa			*Bigelowiella natans*			n	m, p			C
		Retaria	Foraminifera		*Reticulomyxa filosa*				?			
Excavata	Metamonada	Fornicata	Diplomonadida		*Giardia lamblia*	n	n					C
		Parabasalia			*Trichomonas vaginalis*	n						C
	Discoba	Jakobida			*Reclinomonas americana*						m	
		Discicristata	Heterolobosea		*Naegleria gruberi*	n						
			Euglenozoa	Euglenea	*Euglena mutabilis*				?			
				Trypanosomatida	*Trypanosoma brucei*			n[a]	m[a]			C
Relation to supergroups unclear	Apusomonadida				*Thecamonas trahens*	n	n				m	C
	Cryptophyceae				*Guillardia theta*			n	m, p			C
	Haptophyta	Prymnesiophyceae			*Emiliania huxleyi*				m			

n, m, p, and a indicate the identification of sequences in the respective phylogenetic subgroup and their predicted or experimentally verified localization to either the nucleus, mitochondria, plastids or apicoplasts, respectively; (m), the corresponding genes are found in some mitochondrial genomes, but not for all species; ?, nuclear-encoded sequences for which localization predictions could not be obtained; [a], lineages for which RNase P enzymes were experimentally validated. ′ and –, P RNA candidates with some (′) or more severe (–) deviation from the consensus. Empty cells correspond to lineages where RNase P-related sequences could not be found. Gray cells correspond to lineages in which the organelles do not have a genome. Light gray cells correspond to lineages for which nuclear genome sequencing projects are not complete, although partial sequence information is available. Finally, C indicates the correlation between the predicted occurrence of a given type of enzyme and the absence of the other one (RNP or PRORP RNase P) in a specific lineage and/or compartment.

MBE

interactions with phylum-specific proteins that remain to be identified (fig. 1).

Based on these common features, we searched for putative *PRORP* genes in the three domains of life. We confirmed that PRORP proteins are Eukarya specific, exclusively encoded in nuclear genomes and widely distributed, that is, found in four of the five eukaryal supergroups. The full set of putative PRORPs is given in supplementary table S1, Supplementary Material online, and summarized in table 1. Briefly, among Opisthokonta, PRORPs are present in Metazoa and all the associated lineages (Choanomonada, Filasterea, and Ichthyosporea), but absent from fungi and associated lineages. No PRORP could be identified in the supergroup of Amoebozoa. Among Archaeplastida, PRORP was not found in the basal groups such as Glaucophyta and Rhodophyta, but was found in all Chlorophyta and Charophyta as single genes, while in Embryophyta, more than two PRORPs were typically found. In Spermatophyta, PRORP sequences can be subdivided into three evolutionary distinct clusters that we term cluster I, II, and III (supplementary fig. S9, Supplementary Material online). Most of the species have three PRORPs with one representative of each cluster. However, the Brassicaceae (e.g., *Arabidopsis*) make an exception, because *Arabidopsis* PRORP2 and 3 both belong to cluster III. PRORPs are also found in the supergroup SAR, two to three PRORP proteins are encoded in all Stramenopiles. In Alveolata, no genes coding for PRORPs were found in ciliates, but a single gene could be identified in all Apicomplexa genomes. Among Excavata, PRORP is found in the sequenced genomes of some Discoba organisms but not in Metamonada. Although present in Euglenozoa, it is not identifiable in Heterolobosea.

To gain insight into the origin and distribution of PRORP, a phylogenetic analysis was performed. The results suggest an ancient origin of PRORP (supplementary fig. S10, Supplementary Material online). Still, in some instances PRORP might also have spread during horizontal gene transfer (HGT) events such as secondary and tertiary endosymbiosis. This might have happened, for example, in stramenopiles where, among individual species, multiple PRORPs cluster in evolutionary distinct groups (supplementary fig. S10, Supplementary Material online).

Although the prevalence of PRORP in Eukarya could be established, understanding the distribution of RNP and PRORP in specific compartments requires to know the precise subcellular localization of PRORPs in the respective lineages. To gain such information, we applied localization prediction tools to full-length PRORP sequences. The results are compiled in supplementary table S2, Supplementary Material online, and summarized in table 1. In short, in Opisthonkonta, all animal PRORPs are mitochondrial. In green algae single PRORP genes might encode both nuclear and organellar PRORPs expressed by alternative translation starts. In land plants, cluster III contains nuclear orthologs of PRORP, while cluster I and II PRORPs are predicted to be organellar. In other groups, SAR, Excavata, Crypthophyceae, or Haptophyta, multiple PRORPs can be targeted to mt and nuclei, or a single PRORP can be found in specific compartments as, for example, in the apicoplast of apicomplexan. Overall, the predicted localizations confirm that PRORP proteins are not restricted to organelles as initially envisaged (Lai et al. 2010), but demonstrates that they are also widespread in nuclei.

Conclusions and Possible Scenarios for the Evolution of RNase P Distribution

In most instances our analyses revealed a correlation between the predicted occurrence of a given type of enzyme (RNP RNase P or PRORP) and the absence of the other one in a specific lineage and/or compartment. The most divergent examples are fungi, where RNP enzymes are active in both mt and nuclei while PRORP is absent, and Streptophyta or Trypanosomatida, where PRORPs are found in organelles and nuclei, whereas RNP genes are absent. Similar correlations are summarized in table 1 for all Eukarya groups.

Our analysis implies that PRORP might have evolved very early during eukaryal evolution, in an organism at the root of modern Eukarya (fig. 2), although its distribution points to some HGT events as well. It appears likely that the fusion of PPR, NYN, and all the features defining PRORP took place only once during evolution. The RNP and protein-only forms of RNase P thus probably coexisted in an early eukaryote, a functional redundancy that, however, might not have persisted in any organism to the present. We did not find solid evidence for this coexistence within the same compartment, although it cannot be ruled out for some Mamiellophyceae, where isoforms of PRORP might be targeted to both nuclei and organelles while RNP RNase P has been retained in organelles. RNP was kept in some organisms (fungi) or compartment (nucleus of metazoa) and protein-only enzymes were not retained. In these organisms, RNPs might have gained additional functions that could not be provided by PRORP, for example, as observed in human nuclei with the requirement of RNP RNase P for the formation of RNA polymerase III initiation complexes (Serruya et al. 2015). In contrast, PRORP was kept in other organisms (some chlorophytes, streptophytes, trypanosomids) or in specific compartments (nucleus of other chlorophytes and mt of metazoans) and RNPs were lost. Similarly, PRORPs targeted to organelles might have coexisted with RNP RNases P encoded in organellar genomes. P RNA genes might have been lost in the course of rearrangements of organellar genomes, consolidating PRORP as the RNase P enzyme in this compartment.

In animal and plant lineages, RNase P distribution followed two different routes. Unicellular organisms basal to Metazoa (Ichtyosporea, Filasterea, Choanomonada) seem to have retained PRORP proteins for mitochondrial RNase P function and this status was also preserved in all metazoan species. In contrast, unicellular organisms basal to Chlorophyta seem to have initially retained PRORP enzymes only for nuclear RNase P activity. Then, in more recent species of the Chloroplastida lineage, PRORP also took over the organellar RNase P function.

In conclusion, looking at the global picture, since its origin PRORP seems to have been an invasive enzyme, taking over

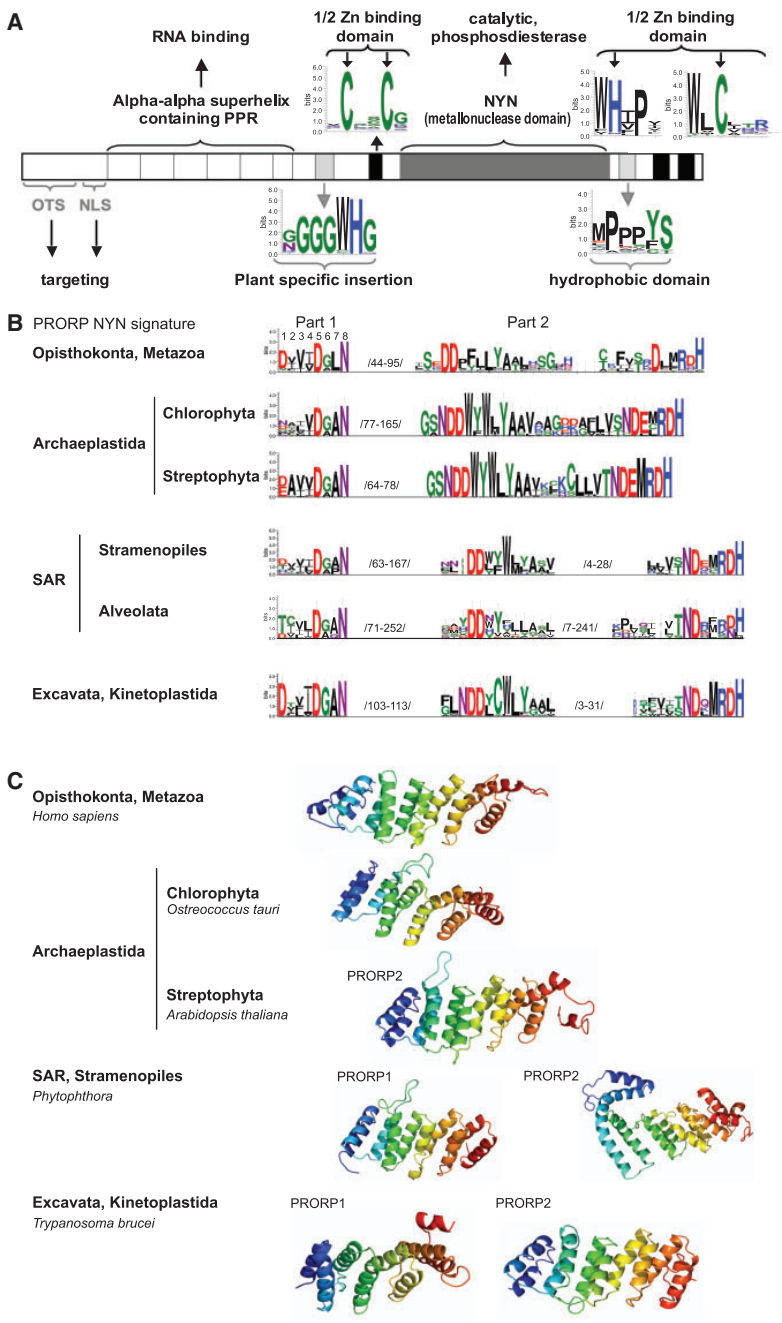


Fig. 1. Description of the conserved features defining PRORP proteins. (*A*) Schematic representation of the different domains of PRORP. Sequence logos of residue conservation for the subdomains involved in zinc binding, as well as for a plant-specific glycine-rich insertion and for a "hydrophobic domain" conserved in organisms that contain a plastid (or had contained a plastid) were generated with WebLogo 3. The number of sequences analyzed and the percentage of sequences originating from animals (Metazoa), plants (land plants), or other organisms (Chlorophyta, Stramenopiles, Alveolata, Cryptophyceae, Haptophyta, Rhizaria, Choanoflagellates, Filastera, Ichtyosporea) are as follows from left to right: Plant-specific insertion: 169 sequences (100% land plants); N-terminal ½ Zn binding domain 1: 275 sequences (1/2 land plants, 1/3 metazoa, 1/6 others); hydrophobic domain: 138 sequences (60% land plants); C-terminal ½ Zn binding domain 2: 249 sequences (1/3 land plant, 1/3 metazoa, 1/3 other). OTS, organellar targeting signals (to mitochondria, plastids, or apicoplasts); NLS, nuclear localization signal. (*B*) Conserved residues present in the PRORP-defining NYN domain signatures, specified for different phyla. The positions of the eight residues constituting part 1 of the NYN signature of PRORP have been numbered as indicated above the first logo. Numbers between the conserved motifs indicate the distance range (in amino acids) that separate the motifs in the different PRORPs analysed. (*C*) Three-dimensional structure predictions for N-terminal domains of representative PRORP proteins considered in this analysis. All the putative PRORPs have an α-superhelical domain consistent with the conserved fold of PPR proteins. N-terminal extremities are shown on the left, C-terminal ones on the right.

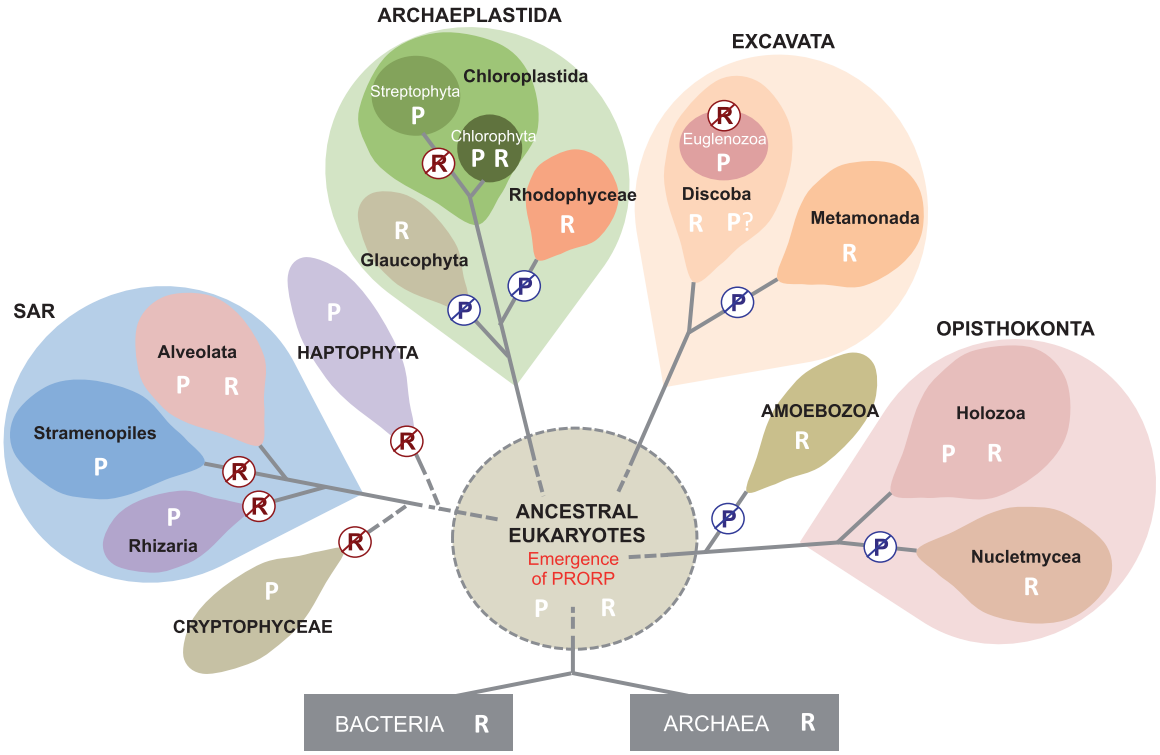


FIG. 2. Distribution of RNP and PRORP RNase P enzymes in the eukaryal domain of life. Relations between eukaryal groups are schematically indicated according to Petersen et al. (2014). R and P indicate the occurrence of RNP and PRORP RNase P enzymes in the respective groups, based on the study presented here. Crossing out P or R indicates putative evolutionary events associated with the loss of PRORP or (nuclear) RNP RNase P. The question mark indicates an example where limited genomic data prevented conclusions as to the occurrence of the given enzyme type in the respective group. The diagram highlights how the distribution of RNase P seemingly involved multiple events of losses of either PRORP or RNP RNase P.

the function of ancestral RNP RNase P in several eukaryal groups, in entire organisms, or in given cellular compartments. The evolutionary trend to replace RNP with PRORP becomes plausible if one considers its capability to instantly replace RNP enzymes in tRNA biogenesis, as experimentally demonstrated for the *E. coli* and yeast systems (Gobert et al. 2010; Taschner et al. 2012; Weber et al. 2014). This evolution may witness a still continuing transitional process from the RNA to the protein world.

Materials and Methods

Identification of Nuclear-Encoded RNase P RNAs

We identified P RNAs using Infernal (Nawrocki and Eddy 2013) with an *E*-value threshold of 1×10^{-8} based on the RFAM 12.0 (Nawrocki et al. 2015) models RF00009 (Nuclear RNase P) and RF01577 (*Plasmodium* RNase P). In addition, we used the tool Bcheck (Yusuf et al. 2010) with default parameters. The predictions were curated and assessed manually for their conserved core. This ensemble of methods also allows discriminating P RNAs from MRP RNAs.

Search for Homologs of the RNP RNase P-Specific Protein Subunit RPP21

We selected reference sequences from several sources: 1) The Rpr2 alignment provided by Rosenblad et al. (2006), 2) the seed alignment provided for the PFAM family PF04032 (RNase P Rpr2/Rpp21/SNM1 subunit domain) (Finn et al. 2011), and 3) WormBase version WS247 (Harris et al. 2010) gene Y37E11B.6 (rpp21). Reference domains were identified and a scoring algorithm was implemented based on regular expressions.

Identification of Rpm2p and Mitochondrial P RNAs in Fungi

The HMMER algorithm (Finn et al. 2011) as well as BLAST searches (Altschul et al. 1990) were used to retrieve proteins with homology to the Rpm2 domain as defined in PFAM (Finn et al. 2014). Putative *rpm1* was retrieved from unannotated fungal mitochondrial genomes with RNAweasel (Gautheret and Lambert 2001).

PRORP Sequence Analysis and Structural Predictions

PRORP sequences were retrieved using the BLAST tool in NCBI (National Center for Biotechnology Information), Ensembl, Bogas, Phytozome, JGI, and Broad. The proteins were aligned using MUSCLE (Edgar 2004). The sequences of these domains were then retrieved and realigned with MUSCLE before using WebLogo 3 (Crooks et al. 2004) to highlight the conserved residues. Protein structures were predicted using the Phyre2 algorithm in the intensive modeling mode (Kelley and Sternberg 2009).

Subcellular Localization Predictions

Subcellular localization predictions were determined for most proteins with TargetP, Predotar, and MultiLoc2 when applicable (Small et al. 2004; Emanuelsson et al. 2007; Blum et al. 2009). PredAlgo was used for PRORP sequences of green algae (Chlorophyta) (Tardif et al. 2012). PlasmoAP and PATS were used for *Apicomplexa* PRORP in order to determine if they possess an apicoplast targeting peptide (Zuegge et al. 2001; Foth et al. 2003).

Phylogenetic Analyses of PRORP

Phylogenetic analysis of PRORP protein sequences were performed with the maximum-likelihood method with 100 bootstrap replicates (Dereeper et al. 2008).

Supplementary Material

Supplementary methods, results, tables S1 and S2, and figures S1–S10 are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

This work was supported by the "Centre National de la Recherche Scientifique," the University of Strasbourg, the Medical University of Vienna, and the Philipps-University of Marburg. We thank Prof. B.F. Lang for critical discussion on the evolution of RNase P. This work was supported by Agence Nationale de la Recherche (grant PRO-RNase P, ANR 11 BSV8 008 01 to P.G.), LabEx consortium "MitoCross," the German Research Foundation (grants HA 1672/17-1 and IRTG 1384 to R.K.H.), and the Austrian Science Fund (grant I299 to W.R).

References

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. *J Mol Biol.* 215:403–410.

Anantharaman V, Aravind L. 2006. The NYN domains: novel predicted RNAses with a PIN domain-like fold. *RNA Biol.* 3:18–27.

Blum T, Briesemeister S, Kohlbacher O. 2009. MultiLoc2: integrating phylogeny and Gene Ontology terms improves subcellular protein localization prediction. *BMC Bioinformatics* 10:274.

Crooks GE, Hon G, Chandonia JM, Brenner SE. 2004. WebLogo: a sequence logo generator. *Genome Res.* 14:1188–1190.

Daoud R, Forget L, Lang BF. 2012. Yeast mitochondrial RNase P, RNase Z and the RNA degradosome are part of a stable supercomplex. *Nucleic Acids Res.* 40:1728–1736.

Dereeper A, Guignon V, Blanc G, Audic S, Buffet S, Chevenet F, Dufayard JF, Guindon S, Lefort V, Lescot M, et al. 2008. Phylogeny.fr: robust phylogenetic analysis for the non-specialist. *Nucleic Acids Res.* 36:W465–W469.

Edgar RC. 2004. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics* 5:113.

Ellis JC, Brown JW. 2010. The evolution of RNase P and its RNA. In: Liu F, Altman S, editors. Ribonuclease P. New York: Springer. p. 17–40.

Emanuelsson O, Brunak S, von Heijne G, Nielsen H. 2007. Locating proteins in the cell using TargetP, SignalP and related tools. *Nat Protoc.* 2:953–971.

Finn RD, Bateman A, Clements J, Coggill P, Eberhardt RY, Eddy SR, Heger A, Hetherington K, Holm L, Mistry J, et al. 2014. Pfam: the protein families database. *Nucleic Acids Res.* 42:D222–D230.

Finn RD, Clements J, Eddy SR. 2011. HMMER web server: interactive sequence similarity searching. *Nucleic Acids Res.* 39:W29–W37.

Foth BJ, Ralph SA, Tonkin CJ, Struck NS, Fraunholz M, Roos DS, Cowman AF, McFadden GI. 2003. Dissecting apicoplast targeting in the malaria parasite *Plasmodium falciparum*. *Science* 299:705–708.

Gautheret D, Lambert A. 2001. Direct RNA motif definition and identification from multiple sequence alignments using secondary structure profiles. *J Mol Biol.* 313:1003–1011.

Gobert A, Gutmann B, Taschner A, Gößringer M, Holzmann J, Hartmann RK, Rossmanith W, Giegé P. 2010. A single *Arabidopsis* organellar protein has RNase P activity. *Nat Struct Mol Biol.* 17:740–744.

Gobert A, Pinker F, Fuchsbauer O, Gutmann B, Boutin R, Roblin P, Sauter C, Giegé P. 2013. Structural insights into protein-only RNase P complexed with tRNA. *Nat Commun.* 4:1353.

Gutmann B, Gobert A, Giegé P. 2012. PRORP proteins support RNase P activity in both organelles and the nucleus in *Arabidopsis*. *Genes Dev.* 26:1022–1027.

Harris TW, Antoshechkin I, Bieri T, Blasiar D, Chan J, Chen WJ, De La Cruz N, Davis P, Duesbury M, Fang R, et al. 2010. WormBase: a comprehensive resource for nematode research. *Nucleic Acids Res.* 38:D463–D467.

Hartmann E, Hartmann RK. 2003. The enigma of ribonuclease P evolution. *Trends Genet.* 19:561–569.

Hartmann RK, Gößringer M, Späth B, Fischer S, Marchfelder A. 2009. The making of tRNAs and more—RNase P and tRNase Z. *Prog Mol Biol Transl Sci.* 85:319–368.

Holzmann J, Frank P, Loffler E, Bennett KL, Gerner C, Rossmanith W. 2008. RNase P without RNA: identification and functional reconstitution of the human mitochondrial tRNA processing enzyme. *Cell* 135:462–474.

Jarrous N, Gopalan V. 2010. Archaeal/eukaryal RNase P: subunits, functions and RNA diversification. *Nucleic Acids Res.* 38:7885–7894.

Kelley LA, Sternberg MJE. 2009. Protein structure prediction on the Web: a case study using the Phyre server. *Nat Protoc.* 4:363–371.

Lai LB, Bernal-Bayard P, Mohannath G, Lai SM, Gopalan V, Vioque A. 2011. A functional RNase P protein subunit of bacterial origin in some eukaryotes. *Mol Genet Genomics.* 286:359–369.

Lai LB, Vioque A, Kirsebom LA, Gopalan V. 2010. Unexpected diversity of RNase P, an ancient tRNA processing enzyme: challenges and prospects. *FEBS Lett.* 584:287–296.

Liu F, Altman S. 2010 Ribonuclease P. New York: Springer.

Marquez SM, Harris JK, Kelley ST, Brown JW, Dawson SC, Roberts EC, Pace NR. 2005. Structural implications of novel diversity in eucaryal RNase P RNA. *RNA* 11:739–751.

Marvin MC, Engelke DR. 2009a. Broadening the mission of an RNA enzyme. *J Cell Biochem.* 108:1244–1251.

Marvin MC, Engelke DR. 2009b. RNase P: increased versatility through protein complexity? *RNA Biol* 6:40–42.

Morales MJ, Dang YL, Lou YC, Sulo P, Martin NC. 1992. A 105-kDa protein is required for yeast mitochondrial RNase P activity. *Proc Natl Acad Sci U S A.* 89:9875–9879.

Nawrocki EP, Burge SW, Bateman A, Daub J, Eberhardt RY, Eddy SR, Floden EW, Gardner PP, Jones TA, Tate J, et al. 2015. Rfam 12.0: updates to the RNA families database. *Nucleic Acids Res.* 43:D130–D137.

Nawrocki EP, Eddy SR. 2013. Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics* 29:2933–2935.

Petersen J, Ludewig AK, Michael V, Bunk B, Jarek M, Baurain D, Brinkmann H. 2014. *Chromera velia*, endosymbioses and the rhodoplex hypothesis—plastid evolution in cryptophytes, alveolates, stramenopiles, and haptophytes (CASH lineages). *Genome Biol Evol.* 6:666–684.

Piccinelli P, Rosenblad MA, Samuelsson T. 2005. Identification and analysis of ribonuclease P and MRP RNA in a broad range of eukaryotes. *Nucleic Acids Res.* 33:4485–4495.

Pinker F, Bonnard G, Gobert A, Gutmann B, Hammani K, Sauter C, Gegenheimer PA, Giegé P. 2013. PPR proteins shed a new light on RNase P biology. *RNA Biol.* 10:1457–1468.

Rosenblad MA, Lopez MD, Piccinelli P, Samuelsson T. 2006. Inventory and analysis of the protein subunits of the ribonucleases P and MRP

provides further evidence of homology between the yeast and human enzymes. *Nucleic Acids Res.* 34:5145–5156.
Rossmanith W. 2012. Of P and Z: mitochondrial tRNA processing enzymes. *Biochim Biophys Acta.* 1819:1017–1026.
Seif E, Cadieux A, Lang BF. 2006. Hybrid *E. coli*—mitochondrial ribonuclease P RNAs are catalytically active. *RNA* 12:1661–1670.
Seif E, Leigh J, Liu Y, Roewer I, Forget L, Lang BF. 2005. Comparative mitochondrial genomics in zygomycetes: bacteria-like RNase P RNAs, mobile elements and a close source of the group I intron invasion in angiosperms. *Nucleic Acids Res.* 33:734–744.
Serruya R, Orlovetskie N, Reiner R, Dehtiar-Zilber Y, Wesolowski D, Altman S, Jarrous N. 2015. Human RNase P ribonucleoprotein is required for formation of initiation complexes of RNA polymerase III. *Nucleic Acids Res.* 43:5442–5450.
Small I, Peeters N, Legeai F, Lurin C. 2004. Predotar: a tool for rapidly screening proteomes for N-terminal targeting sequences. *Proteomics* 4:1581–1590.
Tardif M, Atteia A, Specht M, Cogne G, Rolland N, Brugiere S, Hippler M, Ferro M, Bruley C, Peltier G, et al. 2012. PredAlgo: a new subcellular localization prediction tool dedicated to green algae. *Mol Biol Evol.* 29:3625–3639.
Taschner A, Weber C, Buzet A, Hartmann RK, Hartig A, Rossmanith W. 2012. Nuclear RNase P of *Trypanosoma brucei*: a single protein in place of the multicomponent RNA-protein complex. *Cell Reports* 2:19–25.
Walker SC, Marvin MC, Engelke DR. 2010. Eukaryote RNase P and RNase MRP. In: Liu F, Altman S, editors. Ribonuclease P, protein reviews. Vol. 10. New York: Springer. p. 173–202.
Weber C, Hartig A, Hartmann RK, Rossmanith W. 2014. Playing RNase P evolution: swapping the RNA catalyst for a protein reveals functional uniformity of highly divergent enzyme forms. *PLoS Genet.* 10:e1004506.
Yusuf D, Marz M, Stadler PF, Hofacker IL. 2010. Bcheck: a wrapper tool for detecting RNase P RNA genes. *BMC Genomics* 11:432.
Zuegge J, Ralph S, Schmuker M, McFadden GI, Schneider G. 2001. Deciphering apicoplast targeting signals—feature extraction from nuclear-encoded precursors of *Plasmodium falciparum* apicoplast proteins. *Gene* 280:19–26.

5.4 Sequence and Structural Properties of Circular RNAs in the Brain of Honeybees (Apis mellifera)

Authors: **Clemens Thölken**, Markus Thamm, Christoph Erbacher and Marcus Lechner

Journal: BMC Genomics

Status: under review

Contributions: Planning and analysis of RNA-Seq experiments, PCR validation of candidates, homolgy analysis, functional annotation, characterization of exon-intron structure, methylation analysis, conserved miRNA target prediction, writing the manuscript

RESEARCH

Sequence and structural properties of circular RNAs in the brain of nurse and forager honeybees (Apis mellifera)

Clemens Thölken[1†], Markus Thamm[2†], Christoph Erbacher[2] and Marcus Lechner[1*]

Abstract

Background: The honeybee (*Apis mellifera*) represents a model organism for social insects displaying behavioral plasticity. This is reflected by an age-dependent task allocation. The most protruding tasks are performed by young nurse bees and older forager bees that take care of the brood inside the hive and collect food from outside the hive, respectively. The molecular mechanism leading to the transition from nurse bees to foragers is currently under intense research. Circular RNAs, however, were not considered in this context so far. As of today, this group of non-coding RNAs was only known to exist in two other insects, *Drosophila melanogaster* and *Bombyx mori*. Here we complement the state of circular RNA research with the first characterization in a social insect.

Results: We identified numerous circular RNAs in the brain of *A. mellifera* nurse bees and forager bees using RNA-Seq with exonuclease enrichment. Presence and circularity were verified for the most abundant representatives. Back-splicing in honeybee occurs further towards the end of transcripts and in transcripts with a high number of exons. The occurrence of circularized exons is correlated with length and CpG-content of their flanking introns. The latter coincides with increased DNA-methylation in the respective loci. For two prominent circular RNAs the abundance in worker bee brains was quantified in TaqMan assays. In line with previous findings of circular RNAs in *Drosophila*, circAm*rsmep2* accumulates with increasing age of the insect. In contrast, the levels of circAm*rad* appear age-independent and correlate with the bee's task. Its parental gene is related to amnesia-resistant memory.

Conclusions: We provide the first characterization of circRNAs in a social insect. Many of the RNAs identified here show homologies to circular RNAs found in *Drosophila* and *Bombyx*, indicating that circular RNAs are a common feature among insects. We find that exon circularization is correlated to DNA-methylation at the flanking introns. The levels of circAm*rad* suggest a task-dependent abundance that is decoupled from age. Moreover, a GO term analysis shows an enrichment of task-related functions. We conclude that circular RNAs could be relevant for task allocation in honeybee and should be investigated further in this context.

Keywords: circRNA; circular transcriptome sequencing; honeybee; brain; neuronal; methylation; CpG; alternative splicing; behavioral plasticity

*Correspondence: lechner@staff.uni-marburg.de
[1]Philipps-Universität Marburg, Institut für Pharmazeutische Chemie, Marbacher Weg 6, 35032 Marburg, Germany
Full list of author information is available at the end of the article
[†]Equal contributor

Background

Honeybees (*Apis mellifera*) display a striking behavioral plasticity among their workers that is reflected in an age-dependent task allocation and thus represent a substantial model organism for phenotypic plasticity. Workers are able to execute varying specific behaviors in order to fulfill tasks that are essential for the viability of the colony, such as cleaning combs, feeding the larvae, guarding the nest entrance and foraging for food. To ensure economic efficiency and to prevent randomly performed tasks, the assignment of tasks has to be coordinated [1]. Task allocation is predominantly dependent on the age of the worker bees, but is also flexible and can be adjusted to colony needs [2, 3, 4]. In experimental single cohort colonies (SCCs) that are solely composed of young bees, some colony members initiate foraging precociously irrespective of their age [5]. Major differences in task-related behaviors exist between the typically younger nurse bees that feed the larvae inside the hive and the older foragers ($\geq$18 days after emergence) that leave the hive to collect pollen, nectar and water [6, 7, 8, 9]. This phenotypic plasticity is also reflected at the neuronal level. The overall brain volume is increased in forager bees compared to nurse bees [10, 11] especially in visually innervated brain structures [12, 13]. At synaptic levels, these changes involve for instance the density of synaptic complexes within mushroom body calyces caused by the growth of Kenyon cell dendrites and pruning of presynaptic boutons [14, 15, 16]. The regulation of these processes is poorly understood and seems to be highly complex. Various effectors are known which include the external environment, the colony state and internal stimuli such as (post-)transcriptional changes. Alterations in the expression ratio of hundreds of genes were detected, including those whose gene-products exhibit synaptic functions [17, 18, 19, 20]. Additionally, protein expression is affected as shown for membrane proteome [21] and phosphoproteome changes [22] in the workerbee brain. Task or age-related differences were also observed in the abundance of micro RNAs (miRNAs). Many of the identified miRNAs have a number of putative target genes that also exhibit functions a neural context [23, 24, 25].

Circular RNAs (circRNAs) represent a class of RNA with considerable regulatory potential that was overlooked for decades and is currently under extensive research and discussion. An increasing number of studies show that circRNAs are abundant, differentially expressed and even have biological functions [26, 27]. In general circRNAs arise from a back-splicing event. The 5'-end of a donor exon is joined to a 3'-end of an acceptor exon of the same molecule [28]. This results in a so called back-spliced junction (BSJ) which can be observed as junction-spanning reads (JSR) when mapping RNA-Seq data to a genome using a mapper that supports split reads. The abundance of circRNAs typically varies between tissues and is sometimes uncorrelated to the host mRNAs [29]. This may indicate a specific function of circRNAs but might as well reflect distinct decay rates for linear compared to circular transcripts which lack accessible ends. Studies point out that circRNAs may act as regulators of alternative splicing [30] or could feature miRNA sponges [31, 26].

Besides human and mice, the presence of circRNAs was verified and studied extensively in the fruit fly *Drosophila melanogaster* and very recently in the silkworm *Bombyx mori* but no other insect so far [26, 29, 28, 32]. Important findings are the presence of numerous canonical miRNA seed matches in line with a putative miRNA sponge function as well as the fact that circRNAs mainly derive from neural genes and accumulate in neural tissues in an age-depended manner [33, 30, 34]. Following up on these findings, circular RNAs may contribute in regulating the age-related transition from nurse bees to foragers at the molecular level.

Results and Discussion

Identification of circRNAs in the brain of honeybees

As circRNAs do not feature 5'- or 3'-ends they are virtually resistant to RNase R treatment, which digests most linear RNAs. The enzyme can thus be used to enrich total RNA extracts for circRNAs [35, 36]. In order to identify these, we prepared RNA-Seq libraries from total RNA extracts of honeybee worker brains. The libraries were enriched for circular RNAs and compared to a non-enriched library. Each BSJ was considered as representative of a distinct circular RNA. We were able to detect a total of 3,384 individual BSJs supported by at least three JSRs from the four libraries combining two different methods, see Figure 1A. Based on these we provide two sets of circRNAs identified by applying different stringency thresholds (see Material and Methods for details). The low stringency sets contains 1,263 circRNAs found by both independent algorithmic methods (overlap). Only these BSJs were considered viable circRNA candidates because previous studies showed inconsistent results between different algorithms [37, 38]. Specifically, `segemehl` is known to produce very sensitive mapping results, potentially introducing false positives when solely relied upon. The high stringency set is a subset containing 254 circRNAs with a higher amount of supporting reads along with a significant five-fold enrichment of the JSRs through RNase R treatment. The majority of the circular transcripts were even enriched by more than ten-fold ($>$ 77%). We remark, that these

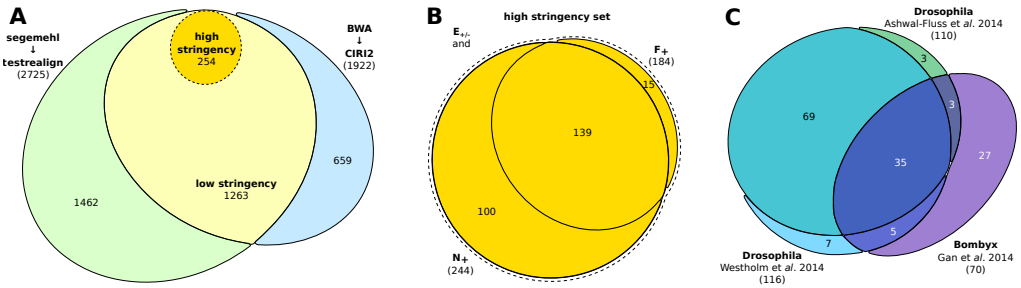


Figure 1 Identified circRNAs by RNA-Seq. (A) Two independent algorithms were used to predict circRNAs: `segemehl` in conjunction with `testrealign` (green) reported 2,725 BSJs while `BWA` with `CIRI2` reported 1,922 BSJs. The overlap was considered as low stringency set (yellow). The high stringency set additionally requires an enrichment through RNase R treatment and compelling read coverage from at least two independent sequencing libraries. See Materials and Methods for details. These 254 candidates are most likely genuine circRNAs and are therefore used for further analysis. (B) Strong evidence is found in three independent sequencing libraries for 139 candidates. The remaining 115 are well supported by two samples. Note that all candidates were also identified as enriched in E_+ vs E_-. (C) 122 host genes are orthologous to host genes of circRNAs identified in either *Drosophila* or *Bombyx* in previous studies.

numbers refer to circRNAs that are expressed in the brain of nurse and forager bees. In contrast, 2,513 circRNAs reported for *D. melanogaster* [33] and 3,916 for *B. mori* [32] are based on samples of different developmental stages, tissues and even cultured cells and do not ensure RNase R enrichment.

Further analyses were performed using the high stringency set. We argue that enrichment control is necessary to discern genuine circRNAs from potential trans-splicing or exon-shuffling events. Otherwise independent experiments would be required to further support the sequencing-based evidence which is not feasible given the high number of involved loci. Focussing on this significantly expressed subset of circRNAs allows us to investigate genomic properties that are an inherent part of circRNA deriving loci. The inclusion of candidates with less confidence (less enrichment, fewer supporting JSRs) would introduce additional noise into statistical analyses. This observations was made e.g. regarding lower read numbers in a *D. melanogaster* study [30].

A vast majority of BSJs was flanked by a canonical GT/AG splice signal. Only five circRNAs did not show such a canonical splice site (see Additional File 1). In one case an annotation was not possible. The BSJ spans two exons that are (presumably) not spliced together. The coding exon of gene CG45167 (homolog of B52 in *D. melanogaster*) and its immediate downstream exon which starts with the 5'-UTR are not present in any currently annotated transcript variant. Details are illustrated in Additional File 1.

The amount of canonically spliced transcripts (linear) is at least the same as the amount of back-spliced transcripts (circular) for the majority of circRNAs identified here. For this reason it is unlikely that the circRNAs presented here arose from a mapping artifact, e.g. due to misalignment of reads or repeating gene copies. We picked some of the most significant circRNAs that were highly abundant or showed a particularly differential expression pattern between nurse or forager bee libraries. Presence and circularity of these selected circRNAs was verified further by additional PCR experiments, see Additional File 6. TaqMan based Real Time PCR assays were used to examine the expression levels in nurse or forager bee for two salient circRNAs in an independent experimental approach, see "Quantification in nurses and foragers". A complete list of all 254 high confidence circRNAs including read levels and putative homologs in *D. melanogaster* and *B. mori* can be found in Additional File 2. An excerpt of the most prominent entities is shown in Table 1.

Homologs to fly and silkworm

Honeybee circRNAs were compared to those found in fruit fly [33, 30] and silkworm [32] based on homology of their parental genes (Figure 1B). Out of 254 honeybee circRNAs only 70 host gene homologs were found in silkworm (30%). In contrast, 203 homologs were identified for fruit fly (80%) which can be explained by the closer phylogenetic relationship to honeybee [40]. Consistent with our results, circularized exons in fruit fly were found in 144-151 of these homologs (with respect to [30] and [33], overlap 122 circRNAs). This finding is in line with a similar comparison of circRNAs in human and mouse. There, two thirds of all host genes harboring back-splicing junctions could be correlated by homologies between the two species [41]. A complete listing of the results can be found in Additional File 2. Even though circRNAs are known for

Table 1 Excerpt of identified circRNAs in the brain of honeybee nurse and forager bees. All circRNAs were significantly enriched in E_+ over the non-enriched set E_-. *Set* refers to the RNA-Seq libraries in which the circRNA was enriched in addition (see Table 2). The *host gene* is given according to the RefSeq GCF_000002195.4 annotation along with the corresponding *BeeBase* identifier [39]. The respective chromosome is indicated in the *Chr.* column. The summarized number of *JSR*s is given along with the averaged normalized expression levels relative to the host gene expression *expr.* and fold enrichment *enriched*. The *homology* column indicates whether a *Drosophila* or *Bombyx* homolog was found in ⋆[30] ○[33] or †[32]. The first block corresponds to circRNAs that were particularly strongly expressed (many JSRs or high norm. expression) or showed signs of differential expression between nurse and forager bee libraries and were thus selected for verification of circularity and presence in further PCR experiments. The full list can be found in the Additional File 2.

circRNA ID	host gene	BeeBase	Chr.	JSRs	enriched	expr.	homology
ame_circ_0001970	LOC413427	GB43145	LG11	432	6.7	0.329	⋆ ○
ame_circ_0000721	LOC724885	GB53835	LG3	139	10.2	0.344	
ame_circ_0002142	LOC410393	GB52063	LG12	124	5.0	0.628	†
ame_circ_0000163	LOC408576	GB42249	LG1	103	7.2	0.328	⋆ ○
ame_circ_0000232	Mup2	GB49259	LG1	97	11.0	0.119	
ame_circ_0001780	rad	GB49511	LG10	91	5.0	0.062	⋆ ○
ame_circ_0001286	LOC411534	GB44365	LG7	63	93.0	0.731	⋆ ○
ame_circ_0002579	LOC409655	GB47584	LG16	42	17.5	0.226	
ame_circ_0001822	Rsmep2	GB54272	LG10	34	5.0	0.022	⋆ ○
ame_circ_0002577	LOC409655	GB47584	LG16	17	11.3	0.072	
ame_circ_0001852	CoRest	GB52614	LG10	10	5.0	0.013	⋆ ○
ame_circ_0001099	LOC411114	GB44582	LG5	306	18.5	0.328	†
ame_circ_0000414	LOC725294	GB55364	LG2	216	7.6	0.312	⋆ ○
ame_circ_0000397	LOC408688	GB49767	LG2	185	5.0	0.377	⋆ ○ †
ame_circ_0001712	LOC408996	GB42579	LG9	169	6.3	0.198	⋆ ○ †
ame_circ_0002576	LOC409655	GB47584	LG16	168	14.9	0.644	
ame_circ_0001638	LOC411347	GB17597	LG9	159	9.4	0.400	
ame_circ_0001593	LOC408991	GB53310	LG9	148	7.9	0.105	
ame_circ_0001479	LOC408957	GB40504	LG8	147	18.4	0.339	○
ame_circ_0000524	LOC408718	GB43446	LG2	130	36.2	0.313	
ame_circ_0001120	sGC-alpha1	GB52929	LG6	129	10.4	0.276	†
ame_circ_0000054	LOC726544	GB42188	LG1	124	7.5	0.480	
ame_circ_0001877	LOC408309	GB45167	LG11	121	9.4	0.085	
ame_circ_0000669	LOC410044	GB55791	LG3	118	13.0	0.370	
ame_circ_0000073	LOC410717	GB55293	LG1	111	11.0	0.290	⋆ ○
ame_circ_0001340	LOC411229	GB42567	LG7	109	6.9	0.570	○

only three insects so far, the number of homologous host genes among them suggests that circRNAs are commonly found in insects. Features identified for circRNAs in one of these organisms are likely to be valid for other insects.

GO term enrichment

A GO term analysis (gene ontology term enrichment) was performed using all 203 circRNA host gene homologs correlated to fruit fly from which we extrapolated the functional annotation. High level processes involved in synaptic development and regulation were significantly enriched. Given that the source samples were obtained from brain tissue, this is an expected result but it also resembles the finding that neurologically associated genes are a main source of circRNAs as found for *D. melanogaster* [33]. The most enriched high level terms below a p-value of 10^{-4} were "anesthesia-resistant, medium- and long-term memory" (27x) "medium-term memory" (23x), "regulation of neuromuscular synaptic transmission" (21x) and "deactivation of rhodopsin mediated signaling" (21x). The former is especially remarkable. One representative of this group is the *radish* gene from which the circRNA circAm*rad* (*ame_circ_0001780*) arises. We found that the abundance levels of circAm*rad* correlated with the acquired task of a bee (see "Abundance and task allocation" below). Consistent with this is also the enrichment of rhodopsin signaling and memory related genes. Nurse bees take care of the brood inside the hive, where it is dark and the requirements to memory are different from those of foragers [42]. After task transition to forager bees, they start to collect food from outside the hive, mostly at daylight, and need to find their way back to the hive afterwards. A need for adaptation of rhodopsin signaling and a change in memory requirements is obvious. In fact, "positive phototaxis" showed the highest GO term enrichment (44x). The significance ($p = 1.87\text{x}10^{-3}$) however was above the applied threshold because the term only has four representatives in the reference set. A detailed overview of enriched GO terms can be found in Additional File 3.

Exon-intron structures

The majority of BSJs in honeybee correspond exactly to exon boundaries of protein coding regions (78%), see Figure 2. Nearly all remaining cases are derived from 5'-UTR containing segments (17%). This is only slightly different from the set of (presumably) linearly spliced exons in the control but shows

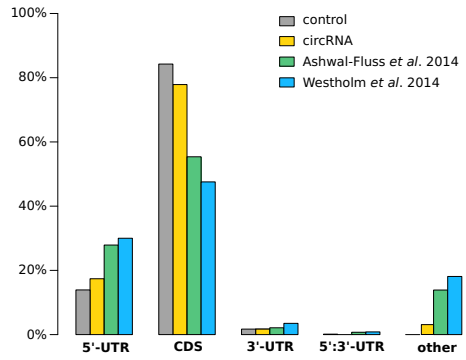


Figure 2 Location of circRNAs in the original transcripts. BSJs are clustered into the following categories: part of the 5'-UTR, from the coding sequence exclusively, part of the 3'-UTR, spanning from 5' to 3'-UTR or other (in the middle of exons, introns, part of non-coding genes or in between genes). Most of the circRNAs in *A. mellifera* originate from exons in the coding portion of transcript, which also represents the largest amount of exons transcriptome-wide. Compared to exons of a random control, circRNAs originate slightly more often from 5'-UTR exons as also described for fruit fly. However, we did not find as many BSJ including exons of the 3'-UTR as reported there. All high confidence circRNAs were mapped to annotated splice sites and we could not detect BSJ with sufficient read numbers outside of annotated regions.

a trend towards 5'-UTRs. For both *D. melanogaster* datasets [30, 33] the overall proportion is similar but with a much stronger bias towards 5'-UTRs ($\sim 30\%$) and non-canonical splice events, e.g. occurring in the middle of introns or exons in between genes ($\sim 20\%$, other). The latter category was rarely found for honeybee circRNAs ($< 2\%$). We note that this difference might be a result of different annotation qualities for honeybee (data from 2018) and fruit fly (data from before 2014) and should thus not be over-interpreted.

For fruit fly it was reported that circRNAs mostly originate from the second exon of a transcript [33]. This is also true for honeybee circRNAs. Figure 3, however, shows that this number is implied by the outstanding abundance of transcripts with only two exons. This is also visible in the randomized control distribution. Compared to this set, the observed starts at exon two are actually less than what would be expected. We identified two factors that correlate with back-splicing: 1) The exon position. 2) The number of exons. The further downstream an exon is located in a transcript and the more exons (and thereby splice-junctions) it exhibits, the more likely circRNAs arise from the transcript.

Another finding from fruit fly indicates that circRNAs with higher normalized expression tend to favor earlier exons than less expressed variants [33]. We reproduced this by partitioning the BSJs according to their normalized expression levels, see Figure 3A. A similar trend is visible in our data. However, the shift towards later exons for less expressed circRNAs (e.g. with expression levels < 0.05) is not as pronounced. Notably, our control exons exhibit a much stronger bias towards the second and third exon for the starts of circular junctions than any of the partitions with circularized exons (almost 60%), especially compared to the control used for fruit fly [33]. An alternative stratification of BSJs by their relative fold change in RNA-seq libraries enriched for circRNAs yields the same results, see Figure 3B. The circRNAs presented here do not involve parts of the 5'-UTR more often than expected but transcripts with unusually long 5'-UTRs appear to be prone to circularization at an increased likelihood.

Intronic features

In honeybee, introns flanking circularized exons are significantly longer than those from linearly spliced exons, see Figure 4A. They can span several thousand bases. This result is in line with findings from fruit fly and human [33, 43]. There in addition, flanking introns showed increased levels of reverse complementarity compared to linearly spliced exons. Reverse complementary regions are thought to enhance the likelihood for base-pairing between the introns. This interaction likely guides back-splicing process [44, 34, 45]. Following up on this assumption, introns were reciprocally scanned for reverse complementary matches at sequence-level using `BLAST` [46], see Figure 4B. While the result shows that introns flanking circularized exons are composed of regions with better complementary (represented by higher bitscores) in general, it is also obvious that complementarity is linked to the length of introns. Higher scores of complementarity matches are likely a result of the fact that introns flanking circularized exons are much longer than those from the control set. The most relevant regions for circularization are probably the end of the 5' flanking and the start of the 3' flanking intron, see Figure 5A for a scheme. Even if the comparison is limited to these regions, the difference in complementary matches cannot explain why some exons are circularized and others are not. The median complementarity is about equal to the control introns that flank linear exons even though the latter show much higher variance especially towards introns with hardly any complementarity, see Figure 5B.

An RNA secondary structure prediction using `RNAfold` [47] was used to investigate potential intron-intron interactions more specifically, see Figure 5C. The difference is more obvious using this method. Co-folded complexes of the control introns exhibit much higher minimum free energy scores (MFE), indicating less base-pairing interaction. The difference is highly significant

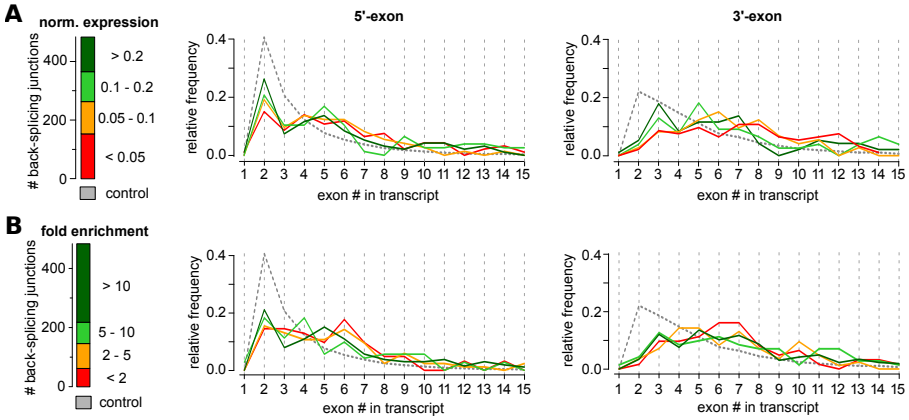


Figure 3 Exon position of circRNAs in the host gene. First exon at the 5'-end the circRNA (left) and last exon at the 3'-end (right) within the host gene, by number of exons. CircRNAs are stratified by (A) normalized circRNA expression into four groups (green to red with decreasing ratio) and (B) relative fold change between RNase R untreated and treated experiments (green to red with decreasing difference). As reference, randomly selected exons in the control (gray) exhibit a higher number of second and third exons in a transcript just by chance.

($p < 0.001$). However, the MFE scores partly cover similar ranges, which does not allow for a clear distinction between circularized exons and linear splicing products. Figure 5D shows that the increase in folding potential (represented by lower MFE scores) is linked to GC-content of the respective introns. Also the fact that the complementarity match as well as the cofolding analysis yielded similar results for all combinations of starts and ends of the flanking introns (e.g. pairing the end of the upstream intron with the end of the downstream intron) puts a direct effect of base-pairing in doubt. The GC-content in turn well discriminates circRNA introns from control introns, see Figure 5E.

Methylation

The intronic features raise the question, why the GC-content of circRNA flanking introns is elevated in such significant amounts (median shifted from 20% to 36%, $p < 0.001$). One reasonable explanation is an increase of potential DNA-methylation at these introns due to CpG islands. While the exact mechanism is unknown so far, DNA-methylation is known to induce alternative splicing in honeybees [48, 49]. Methylation patterns also vary depending on the age and allocated task of an individual bee [50, 51, 52]. It was even shown that reverted nurse bees regain their original methylation patterns independent of their age [53]. Figure 5F illustrates that the CpG dinucleotide frequency is also significantly increased for circRNA flanking introns and nearly absent in the control group ($\sim 1\%$). As CpG sites are preferentially methylated [50, 51], this indicates a significant increase of potential DNA-methylation sites. Moreover, cytosine methylation and hydroxymethylation at non-CG sites (CA, CT, CC) is reported to be enriched in introns of the honeybee [54]. In line with this, Figure 5F shows that also the cytosine mononucleotide frequency is significantly increased for circRNA flanking introns. While the genome comprises $\sim 16\%$ cytosines, circRNA introns exhibit a median of $\sim 18\%$ cytosines. Strikingly, the median cytosine-content of linearized exons is as low as 10%. This can be translated into reduced methylation and hydroxymethylation potential and thereby fewer alternative splicing events for introns flanking canonically spliced RNAs compared to those that frequently result in circRNAs.

We evaluated publicly available whole genome bisulfite sequencing data of worker bees from a previous study to comprehensively determine methylation levels [53]. Figure 5H shows that the length-normalized accumulative DNA-methylation of introns flanking circular RNAs actually tends to be increased compared to those flanking random exons. Notably, the effect was not visible using only the closest 50 or 100 nucleotides of a flanking intron but became visible using a 200 nt window or full-length introns. This is probably due to the limited windows size which is likely too small for statistical assessment.

While relevant social roles in the used methylation study [53] are the same, we note that collection times and extraction methods differ from experiments done in this study. Ideally, the libraries used for circRNA detection and DNA-methylation analysis should be derived from the same biological sample. Without further experimental investigation a strong conclusion

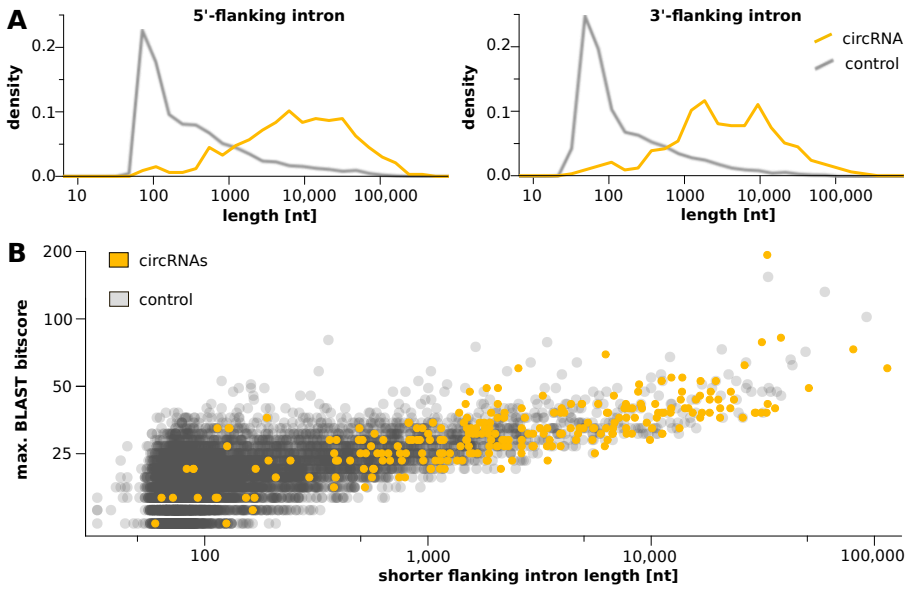


Figure 4 Introns flanking circRNAs. (A) Flanking introns upstream (left) and downstream (right) of circRNAs (yellow) are significantly longer (t-test with $p < 0.001$ for both) than those of a random control (gray). (B) BLAST bitscores of the best reverse complementary match of the shorter circRNA flanking intron to the other (yellow) compared to control intron pairs (gray). Introns flanking circRNAs are visibly longer than random introns, but show similar bitscores for reverse complementary matches of equal intron lengths (correlation $\rho = 0.5$). Notably, shorter introns in the control frequently exhibit relatively high bitscores, indicating that short complementarity regions are common among intron pairs.

cannot be drawn yet. We argue however, that the data presented here provides first indications for a link of circularization and DNA-methylation in honeybees. On this basis we speculate that the age-depended increase of circRNA abundance is not (only) due to potentially lower decay rates of circRNAs compared to linear products but also a result of increasing DNA-methylation that leads to alternative splicing accompanied by increase of circRNA formation.

miRNA targets

Potential miRNA target-sites were annotated for all 254 circRNAs identified here. The results can be divided based on their degree of phylogenetic conservation. 3,058 target sites were only conserved in *Apis* species. We argue that *Apis* species are too closely related to qualify as reliable predictor for miRNA target sites. The sequence conservation in this set appears rather high in general. This is also reflected by a similar distribution of potential miRNA target sites compared to the control without any constraints on conservation, see Figure 6A.

A set of 1,076 sites is conserved in *Apis* and eusocial insects which are sufficiently distant to *A. mellifera* to reasonably infer conservation. With about 10.4 target sites per $1,000nt$ circRNAs have a 1.7x increase in conserved, putative miRNA target sites compared to the median of linear splice product control. Thus, in line with previous findings for *Drosophila* [33], we report a general enrichment of conserved miRNA target sites in circRNAs over random linear counterparts. The most enriched miRNA target sites correspond to ame-miR-3748/ame-miR-3753 ($\sim$ 10x enriched, same seed region) and ame-miR-3791 ($\sim$ 9.2x enriched), see Figure 6B. RNA expression studies show that the abundance levels of some miRNAs correlate with task or age of honeybees [25, 23, 55, 24]. We did, however, not find a significant overlap of miRNAs corresponding to enriched target sites and miRNAs reported as differentially expressed in nurses and foragers. The complete list of potential target sites and their degree of conservation can be found in Additional File 4.

Quantification in nurses and foragers

The circRNAs *ame_circ_0001780* and *ame_circ_0001822* showed a notable differential expression pattern in RNA-Seq results of nurse bees and foragers. For simplicity they will be termed according to their host genes in the further course of the study: circAm*rad* and circAm*rsmep2*, respectively. As the experimental setup is not suitable for any reliable quantitative

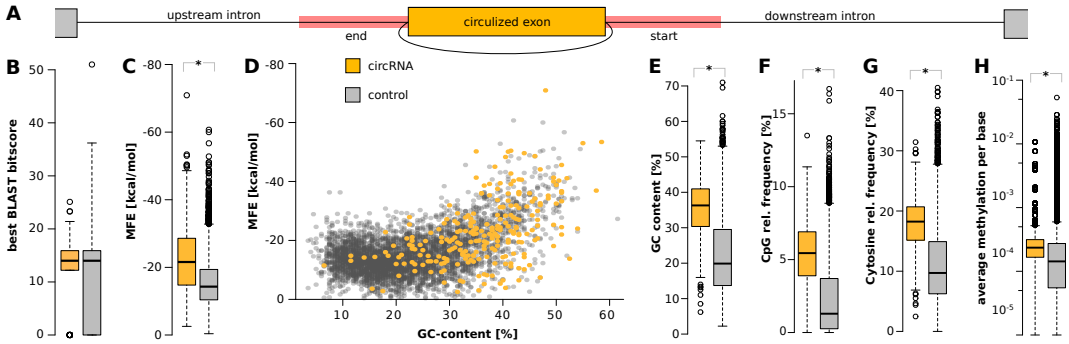


Figure 5 Putative interactions and sequence properties of flanking introns. (A) Scheme indicating the relevant regions (red) of introns flanking circularized exons (yellow). (B) Screen of sequence complementarity using `BLAST`. The scores for circular and linear control intron pairs largely overlap. (C) Screen of intron interactions represented by the MFE of constrained `RNAcofold` calculations. The differences of both groups are significant. Still there is some overlap. (D) Relation of GC-content and MFEs. Both measures appear to be linked (correlation $\rho = 0.5$, Spearman's rank correlation). (E) Screen of GC-content. A significant increase is found compared to the control. (F) Screen of CpG dinucleotides. A significant increase is found compared to the control which rarely exhibits CpG. (G) Screen of cytosine-content. A significant increase is found compared to the control. It explains the observed effects for GC-content and CpG frequency. (H) Screen of average base methylation. The number of methylated bases in introns flanking circRNAs is slightly but significantly increased compared to linear control introns. *p-value < 0.001

assertions, we decided to perform a targeted quantitative Real Time PCR for these candidates at different developmental stages. In addition, we compared the expression patterns in bees with age-related task allocation to those undergoing a task allocation due to colony needs (same-age, SCC), see Figure 7.

For circAm*rsmep2* we found that expression in the brain is higher in foragers than in nurse bees (Figure 7A). This difference, however, does not seem to be directly task-related. In a SCC where nurse and forager bees have exactly the same age, no expression differences are observed (Figure 7A). Our interpretation is that this expression difference most likely depends on the bees' age but not on its task. Supposedly, circAm*rsmep2* accumulates over time in the brain of worker bees, as shown for certain circRNAs in the nervous system from mammals to flies [33, 56]. On the other hand, a significant increase of the linear product in foragers was reported previously (XM_393489.3/Am*rsmep2*, $log_2\ ratio \sim 2.8$) [18]. The observed increase of the circular product circAm*rsmep2* might thus be a consequence of generally increased expression of the host gene, which codes for a RIM-family (Rab3a-interacting molecule) protein. Studies in species of the tetrapoda clade (human, mouse, chicken and so on) show that this family plays an important role in neuronal plasticity, especially in neurotransmitter release and in organizing active zones in plasma membranes [57, 58].

In contrast, circAm*rad* is higher expressed in brains of nurse bees than in brains of foragers (Figure 7B, typical). Strikingly, this is inversely correlated with the expression of the linear product which is strongly increased in foragers (XM_393494.2/Am*rad*, $log_2\ ratio \sim 6.1$) [18] and holds true independent of the age-related task transition. The expression levels in the SCC experiment (Figure 7B, single cohort) are similar to that of typical colonies where tasks are allocated based on a bee's age. This data suggests a correlation of acquired task and circAm*rad* levels. Either the task of the bee is influencing circAm*rad* expression or *vice versa*. Its host gene is orthologous to the *radish* gene in *D. melanogaster*, which is known to play a crucial role in the amnesia-resistant memory (ARM). Unlike the long term memory ARM does not require protein *de novo* synthesis [59] and thus represents a low costs memory form [60, 61]. *Rad* also exhibits circRNAs in fly (Table 1), but whether this circRNA is involved in ARM or whether ARM is also present in honeybees, has not yet been investigated.

Conclusion

1,263 circular RNAs were identified in the brain of honeybees (*A. mellifera*) using RNA-Seq. Adding constraints with respect to read coverage and RNase R enrichment, we yield a set of 254 high confidence circRNAs for further targeted studies. Besides *D. melanogaster* and *B. mori*, this is the third insect for which the existence of circRNAs was shown. Given 1) the evolutionary distance of the three species, 2) the fact that Hymenoptera (*A. mellifera*) branch very early in evolution of holometabolous insects [40], 3) the high amount of homologous circRNAs host genes among

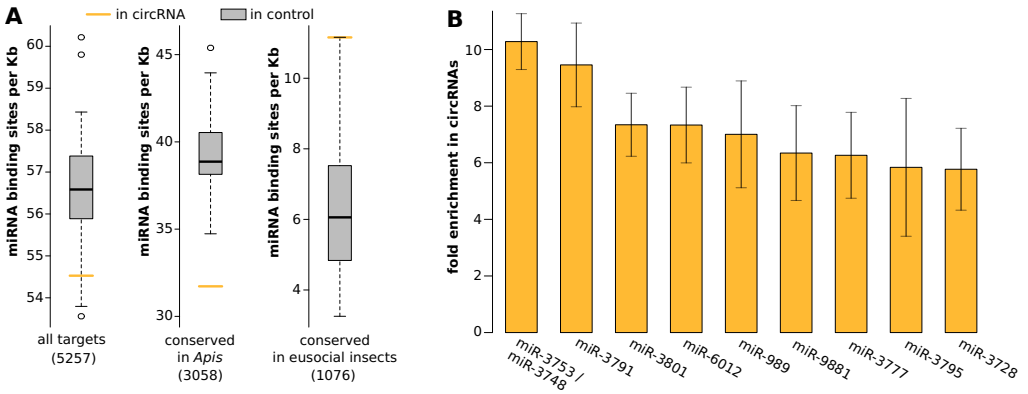


Figure 6 Occurrence of putative miRNA target sites in circRNA exons. (A) Putative target sites normalized by exon length in differentially conserved sets. "All targets" represents no conservation at all. *Apis* represents conservation only in closely related species. Eusocial insects are sufficiently distant to consider conservation relevant. The number of miRNA binding sites conserved in this set is significantly higher ($p < 0.001$, Students t-test) than in the control. The absolute number of potential binding sites in the respective sets is given in parentheses below. (B) Illustration of miRNAs with potential target sites in circRNAs conserved even in eusocial insects. Only miRNAs with at least ten target sites and an at least 5-fold enrichment over control are shown. Error bars indicate standard deviation in fold enrichment between different controls (N=42).

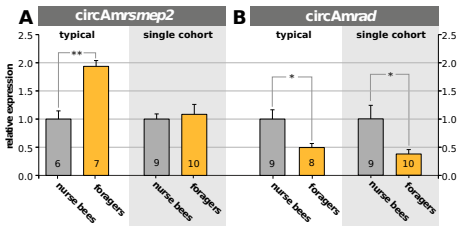


Figure 7 Expression analysis. TaqMen expression analysis of (A) circAm*rsmep2* and (B) circAm*rad* in brains of nurse bees and foragers from colonies with a typical age structure and SCCs consisting of bees of the same age. Expression is given relative to nurse bees. Bars show relative mean expression with standard error. The number of replicates are indicated in the bars. Significant differences are indicated (*$p < 0.05$, **$p < 0.01$, Two tailed unpaired Students t-test). In contrast to circAm*rsmep2* the expression of circAm*rad* seems correlated with the allocated task rather than the bees' age.

these species, and 4) the number of conserved, putative miRNA targets in regions homologous to honeybee circRNA exons in eusocial insects (1076) it can be assumed that circRNAs are a common feature among insects. In line with previous findings from *D. melanogaster* [33] we find a general enrichment of conserved miRNA target sites in circRNAs over random linear counterparts.

Back-splicing in honeybee occurs preferentially towards the end of transcripts and in transcripts with a high number of exons. As reported for *D. melanogaster*, back-splicing is correlated with the length of the 5' and 3' flanking introns [33]. Additionally, a correlation was found regarding the cofolding probability of these intronic regions as well as their CpG- and cytosine-contents which might be relevant for DNA-methylation. In fact, the methylation was found to be increased for circRNA-flanking introns.

A number of circRNAs identified here were confirmed in independent PCR experiments. For two circRNAs, we were able to reliably show a differential expression in brains of nurse and forager bees. The abundance of circAm*rsmep2* (*ame_circ_0001822*) seems to accumulate with the age of the worker bee. This observation is similar to those of circRNAs found in neural tissue of e.g. *Drosophila* [33, 34]. Surprisingly, the expression of circAm*rad* (*ame_circ_0001780*) did not show this accumulation with age but seems to be expressed in a task-dependent manner. Its concentration is found to be reduced in foragers compared to nurse bees in colonies with a typical age-structure as well as in SCCs. This finding is the first indication of a link between the circRNA and the social role of honeybees, which is also indicated in general by the GO term analysis of circRNA host genes. The highest GO term enrichment was found for rhodopsin signaling, phototaxis and memory related genes.

Studying circRNAs in the context of synaptic plasticity and neuronal processes promises further insights into the mechanism of task allocation and behavioral regulation of honeybees and probably also of other insects. New evolving techniques such as genome editing using CRISPR/Cas9 which is also available in honeybees [62] and the micro-injection of short interfering RNAs into the *medial ocellus* [63] will be promising

approaches to study the physiological and behavioral effects of altered circRNA levels. The latter could be used to decrease circRNA levels in the brain by specifically targeting circular junctions and thereby promoting their decay. Genome editing on the other hand might provide means to induce changes to introns that alter the formation of circRNAs.

Methods

Collection of bees

Bees were derived from colonies with normal age structure and with a naturally mated queen located on the grounds of the University of Würzburg. Bees were considered as nurse bees, if they clearly poked their head into open brood cells containing young larvae. Foragers were captured when returning from a foraging flight and having huge pollen loads at their hind legs. Collected bees were frozen in liquid nitrogen immediately. A single cohort colony was established by transferring 2,500 newly emerged bees (marked by the same color immediately after hatching) into a small hive together with one queen in one brood frame and one frame with pollen and honey. Single cohort colony bees were collected at the age of eleven days and controlled for their social task.

RNA-Seq

We used a total of four RNA-Seq libraries to determine circular transcripts present in the brain of honeybees. First, an enrichment control was compiled from the brains of ten dissected nurse bees and ten dissected foragers. Total RNA was extracted with IsolRNA lysis reagent (5PRIME, Hilden, Germany) and treated with DNase I. The sample was divided into two halves. One half (E_+) was treated with 3 units RNase R (epicentre, Madison, USA) per μg total RNA. Digestion was performed for 30min at 37°C. For the other half (E_-) an equivalent volume of double distilled water was added. Afterwards, both samples were purified using phenol–chloroform extraction. Efficacy of the RNase R treatment was verified in a control experiment shown in Additional File 7. Second, we took additional samples from ten nurses and ten foragers separately and treated both with RNase R as described above (samples F_+ and N_+ , respectively) in order to distinguish task dependent expression levels. Library preparation and Illumina© sequencing ($125nt$ paired-end) were performed by GATC Biotech AG (Konstanz, Germany). All RNA-Sequencing data was made publicly available via bioproject PRJNA345404, see Table 2.

Table 2 Summary of RNA-Seq libraries published along with this study. Samples were taken from brains of nurse bees, forager bees or a blend of both.

Sample	SRA ID	Role	Treatment	# Reads
E	SRR4343845	both	-	8,432,479
E_+	SRR4343846	both	RNase R	7,690,777
N_+	SRR4343847	nurse	RNase R	5,843,829
F_+	SRR4343848	forager	RNase R	5,931,097

Identification of circular RNAs

We used two independent algorithmic approaches for the identification of circular RNAs. In one approach reads were mapped to the NCBI *A. mellifera* genome version 4.5 release 102 (RefSeq GCF_000002195.4) using `segemehl` (v0.2.0) with the split reads option [64]. The alignment was subsequently screened for model-free splicing events using the accompanied `testrealign` tool. In the second approach we used `BWA` (v0.7.5a) as mapping tool and subsequently screened using `CIRI2` (v2.0.6) with default parameters [65, 66]. Identified junctions were post-processed using custom scripts bundled in our `Chiasm` suite. `Chiasm` was also used to perform the statistical calculations later on (e.g. CpG-content, pairing-probability, see below). The full analysis pipeline is publicly available at `https://git.io/chiasm`. More precisely, junctions with almost identical start and end positions were merged if they differed by less than 6 nt. Junctions mapped +/-5 nt next to exon boundaries were corrected to exactly match the boundary. This accounts for small variations in sequencing and mapping, e.g. due to flanking intron sequence being potentially identical to the junctioning exon or indels in the genome. We assigned the respective gene and exon numbers to each hit and normalized the number of JSR to the host gene's total read number. Analogously to present studies in *Drosophila* [33] we normalized BSJ read counts ($norm(n_\circ)$) by dividing the number or circular JSRs ($n_\circ$) by the number of mapped library reads N (in millions), divided by reads per kilobases in million reads (RPKM) of the host gene (g). The latter is defined as number or reads assigned to the host gene (n_g) divided by the length of the gene (l_g) in thousand bases and divided by library size of mapped reads, N, in millions.

$$norm(n_\circ) = \frac{n_\circ}{\frac{N}{1{,}000{,}000}\ RPKM_g} \tag{1}$$

with

$$RPKM_g = \frac{n_g}{\frac{l_g}{1{,}000}\ \frac{N}{1{,}000{,}000}} \tag{2}$$

We divided the identified circular RNAs into two sets limited by different stringency levels. The low stringency set contains all circRNAs picked up by both approaches (`testrealign` and `CIRI2`) with at least three JSRs. In the high stringency set, we only considered BSJs with more than ten JSRs across all libraries as

suggested in literature [33]. Thereby, the BSJ has to be found in library E_+ and at least one other independent RNase R treated library. Moreover, a five fold enrichment of JSRs in the RNase R treated library (E_+ vs E_-) is required.

Validation of circRNAs

Total RNA was extracted from ten worker bee brains and prepared as described for the RNA-Seq preparation (see above, without enrichment by RNase R). After DNA digestion, 1μg of RNA were transcribed into cDNA using RevertAid H minus reverse transcriptase (ThermoFisher Scientific) adhering to the manufacturer's specifications. For PCR amplification 15μmol of divergent primers were added to 10ng of cDNA with 25μL of Phusion Polymerase master mix. PCR steps were 30sec heating to 98°C followed by 35 cycles of 10sec denaturation at 98°C, 10sec annealing at 62°C and 8sec elongation at 72°C. After a final extension period of 10min at 72°C, PCR products were either stored at −20°C or subjected to agarose gel electrophoresis. Primer sequences are provided in Additional File 5. The results of PCR verification are provided in Additional File 6.

Quantification of circRNAs

750μL of Isol-RNA lysis reagent (5PRIME, Hilden, Germany) was added to frozen brain samples and homogenized subsequently. After adding 150μL of chloroform and consequent phase separation the aqueous phase was transferred to 900μL ethanol (75%). RNA was purified using peqGOLD Total RNA Kit (Peqlab, Erlangen, Germany) following the standard protocol provided by the manufacturer including an optional DNase I digestion step. From each bee 1.5μg of total brain RNA was transcribed using qScriber cDNA Synthesis Kit (highQu, Kraichtal, Germany). Triplicates of each cDNA (5μL) were run in a quantitative Real Time PCR on a Rotor-Gene Q (Qiagen, Hilden, Germany) in a total reaction volume of 25μL, containing each primer (0.25μM), TaqMan probe (0.1μM), Rotor-Gene Multiplex PCR 9Master Mix (Qiagen, Hilden, Germany). TaqMan probe sequences are provided in Additional File 5. The following protocol was used: 60°C for 1min, 95°C for 5min and 45 cycles at 95°C for 20s and 60°C for 1min. Afterwards the relative expression to AmEF1α [67] with the $\Delta\Delta C_t$ method was determined using Rotor-Gene Q software (Qiagen, Chatsworth, CA). Expression of circRNA was compared only, if respective groups did not differ in their AmEF1α expression ($p > 0.05$, Student's t-test). For the circRNA candidates circAm*rsmep2* and circAm*rad* the established TaqMan probe based assays were designed using outward facing primers. PCR experiments for detection of circRNAs were designed analogously to [28]. The TaqMan probe binds directly to the circular junction and thus signals can only derive from non-canonical spliced RNAs.

Homology screen and functional annotation

Predicted circRNAs were correlated to those previously reported for *D. melanogaster* [33, 30] and *B. mori* [32]. We matched the loci based on the predicted homologs of the closest protein-coding gene with respect to `OrthoDB` v9 [68]. CircRNAs from genes without homolog could thus not be accounted for. Homologous fruit fly genes were then submitted to the online `PANTHER` annotation platform for further over-representation analysis using Fisher's Exact test with false discovery rate (FDR) multiple testing correction. We included functional annotations with more than 5-fold over-representation and FDR below 1%.

Sequence and structural analysis

Based on the genomic annotation and the largest spanning transcript of each circRNA that contained exon boundaries, we extracted whether the circRNA contained part of the 5'-UTR, 3'-UTR of a canonical protein-coding transcript or if it exclusively contained coding regions. The number of exons spanned by the transcript was noted for the 5' and 3' end of the BSJ. For comparison to potentially non-circular transcripts a random control was generated by drawing genes with more than two exons proportionally from all chromosomes and picking exon boundary pairs that were neither from the start nor the end of the transcript. Genes harboring any JSRs found in this study were excluded from this control, see Additional File 2. A random control of 10,000 such junctions were generated for all following statistical tests. Flanking introns were determined by including the sequence outside of the BSJ until the next exon in the same transcript.

In order to screen for complementarity between flanking intron pairs, the 5' intron was matched to the 3' intron using `BLAST` [46] with a word size of six to determine the highest scoring stretch of reverse complementarity. We repeated the procedure with 100nt from the end of the upstream and 100nt from the start of the downstream intron, to discern whether approximate regions showed increased complementarity. The same 100nt portions were used for structural analysis utilizing `RNAcofold` [47]. We applied soft constrains to ensure MFE scores solely based on base-pairing between both intronic regions. Both procedures were repeated with all combinations of starts and ends of the respective introns as educated control set (an interaction of the end of the upstream and the end of the downstream intron is probably not relevant). Surprisingly,

the results for all combinations were similar. To rule out, that we bias for specific length effects at $100nt$, all calculations were also done with 50 and $200nt$ without changing the outcome (data not shown). Introns were checked for GC-content ignoring undetermined residues in the genome sequence (N). Similarly the mononucleotide frequency of cytosine and the relative frequency of CpG dinucleotides was calculated.

To asses whether the observed increase of potential DNA-methylation sites is reflected in actual DNA-methylation, we used whole genome bisulfite sequencing data of worker bees that was publicly available. Precisely, we used all native worker libraries provided in BioProject PRJNA104931 [53] and combined them for this analysis as no differences in average methylation was found between nurse and forager bee libraries for the genes relevant in this study (data not shown). Methylation patterns were analyzed using `Bismark` [69] v0.19.1 with `Bowtie2` [70] v2.2.6 for bisulfite specific mapping and default parameters suggested by its authors. For each intron we counted the average methylation per base on both strands. We required an average coverage of at least five reads for each intron. Calculations were done for 50, 100 and $200nt$ as well as for the length of the complete intron where it exceeded $200nt$ and numbers were normalized by the respective sequence length. A single-sided Wilcoxon-Mann-Whitney rank-sum test was used to determine significance of the increase over the control.

miRNA interference analysis

Predicted and experimentally verified miRNA sequences of *A. mellifera* were obtained from miRBase [71] release 21. Potential target sites were screened in all exon sequences overlapping with the identified circRNAs using nucleotide two to seven of the mature miRNA sequence, see [72]. The analysis pipeline is publicly available on git.io, see above. For each potential miRNA binding site, we determined conservation in further *Apis* species (*A. cerana*, *A. dorsata*, *A. florea*) and other eusocial insects (*E. dilemma*, *L. ventralis*, *M. quadrifasciata*, *B. impatiens*, *B. terrestris*) for the seed region with $100nt$ up- and downstream using the best `BLAST` match [46] in the respective genome. We considered a site conserved if the $6nt$ seed region was perfectly conserved among three out of four *Apis* or four out of five eusocial insects, respectively. As random control we used linear exons, see "Sequence and structural analysis". We split the control to sets of about equal size (42 sets) and applied the above procedure to each set. This results in 42 control datasets where each represents a subset of exons with similar length to avoid a bias due to an over-representation of certain length species. Identified target sites were normalized to sites per $1,000nt$.

Abbreviations

BSJ: back-spliced junction, **circRNA**: circular RNA, **cDNA**: complementary DNA, $\mathbf{C}_t$: threshold cycle, **FDR**: false discovery rate, **GO:** gene ontology, **ID:** identifier, **JSR:** junction-spanning read, **MFE:** minimum free energy, **miRNA:** micro RNA, **PCR:** polymerase chain reaction, **RNA-Seq:** RNA-Sequencing, **SCC:** single cohort colony, **SRA:** sequence read archive, **UTR:** untranslated region

Declarations

Availability of data and materials

All RNA-Seq datasets used in this study are available via NCBI BioProject `PRJNA345404` (https://www.ncbi.nlm.nih.gov/bioproject/PRJNA345404). For details on individual SRA IDs see Table 2. The analysis pipeline is publicly available at https://git.io/chiasm.

Competing interests

The authors declare that they have no competing interests.

Funding

This work was partially funded by the PostDoc Plus Grant to MT of the Graduate School of Life Sciences, University of Würzburg.

Author's contributions

ML and MT conceived the study. ML and CT advised on the sequencing procedure. CE grew bees and conducted the RNA-seq preparation and experiments. CT carried out the bioinformatic analysis. CT and CE performed PCR verifications of circular transcripts. MT carried out the quantification assays. ML, MT and CT wrote the manuscript. All authors reviewed the final manuscript.

Acknowledgements

We thank Marietta Thüring for fruitful discussions.

Author details

[1]Philipps-Universität Marburg, Institut für Pharmazeutische Chemie, Marbacher Weg 6, 35032 Marburg, Germany.
[2]Julius-Maximilians-Universität Würzburg, Verhaltensphysiologie und Soziobiologie, Am Hubland, 97074 Würzburg, Germany.

References

1. Oster, G.F., Wilson, E.O.: Caste and ecology in the social insects. Monographs in population biology **12**, 1–352 (1978)
2. Dreller, C., Page Jr., E.R., Fondrk, K.M.: Regulation of pollen foraging in honeybee colonies: effects of young brood, stored pollen, and empty space. Behavioral Ecology and Sociobiology **45**(3), 227–233 (1999). doi:10.1007/s002650050557
3. Le Conte, Y., Mohammedi, A., Robinson, G.E.: Primer effects of a brood pheromone on honeybee behavioural development. Proceedings. Biological sciences **268**(1463), 163–168 (2001). doi:10.1098/rspb.2000.1345
4. Vaughan, M.D., Calderone, W.N.: Assessment of pollen stores by foragers in colonies of the honey bee, Apis mellifera l. Insectes Sociaux **49**(1), 23–27 (2002). doi:10.1007/s00040-002-8273-3
5. Robinson, G.E., Page, R.E., Strambi, C., Strambi, A.: Hormonal and genetic control of behavioral integration in honey bee colonies. Science (New York, N.Y.) **246**(4926), 109–112 (1989). doi:10.1126/science.246.4926.109
6. Seeley, T.D.: The Wisdom of the Hive. Harvard University Press, Cambridge Mass, London (1995)
7. Thamm, M., Scheiner, R.: Pkg in honey bees: spatial expression, amfor gene expression, sucrose responsiveness, and division of labor. The Journal of comparative neurology **522**(8), 1786–1799 (2014). doi:10.1002/cne.23500
8. Scheiner, R., Reim, T., Søvik, E., Entler, B.V., Barron, A.B., Thamm, M.: Learning, gustatory responsiveness and tyramine differences across nurse and forager honeybees. Journal of Experimental Biology **220**, 1443–1450 (2017). doi:10.1242/jeb.152496
9. Değirmenci, L., Thamm, M., Scheiner, R.: Responses to sugar and sugar receptor gene expression in different social roles of the honeybee (Apis mellifera). Journal of insect physiology (2017). doi:10.1016/j.jinsphys.2017.09.009

10. Withers, G.S., Fahrbach, S.E., Robinson, G.E.: Selective neuroanatomical plasticity and division of labour in the honeybee. Nature **364**(6434), 238–240 (1993). doi:10.1038/364238a0
11. Fahrbach, S.E., Moore, D., Capaldi, E.A., Farris, S.M., Robinson, G.E.: Experience-expectant plasticity in the mushroom bodies of the honeybee. Learning & memory (Cold Spring Harbor, N.Y.) **5**(1-2), 115–123 (1998)
12. Durst, C., Eichmüller, S., Menzel, R.: Development and experience lead to increased volume of subcompartments of the honeybee mushroom body. Behavioral and neural biology **62**(3), 259–263 (1994)
13. Scholl, C., Wang, Y., Krischke, M., Mueller, M.J., Amdam, G.V., Rößler, W.: Light exposure leads to reorganization of microglomeruli in the mushroom bodies and influences juvenile hormone levels in the honeybee. Developmental Neurobiology **74**(11), 1141–1153 (2014). doi:10.1002/dneu.22195
14. Farris, S.M., Robinson, G.E., Fahrbach, S.E.: Experience- and age-related outgrowth of intrinsic neurons in the mushroom bodies of the adult worker honeybee. Journal of Neuroscience **21**(16), 6395–6404 (2001)
15. Groh, C., Ahrens, D., Rössler, W.: Environment- and age-dependent plasticity of synaptic complexes in the mushroom bodies of honeybee queens. Brain, behavior and evolution **68**(1), 1–14 (2006). doi:10.1159/000092309
16. Muenz, T.S., Groh, C., Maisonnasse, A., Le Conte, Y., Plettner, E., Rössler, W.: Neuronal plasticity in the mushroom body calyx during adult maturation in the honeybee and possible pheromonal influences. Dev Neurobiol **75**(12), 1368–1384 (2015). doi:10.1002/dneu.22290
17. Kucharski, R., Maleszka, R.: Evaluation of differential gene expression during behavioral development in the honeybee using microarrays and northern blots. Genome biology **3**, 0007 (2002)
18. Liu, F., Li, W., Li, Z., Zhang, S., Chen, S., Su, S.: High-abundance mRNAs in Apis mellifera: comparison between nurses and foragers. J Insect Physiol **57**(2), 274–279 (2011). doi:10.1016/j.jinsphys.2010.11.015
19. Lutz, C.C., Rodriguez-Zas, S.L., Fahrbach, S.E., Robinson, G.E.: Transcriptional response to foraging experience in the honey bee mushroom bodies. Dev Neurobiol **72**(2), 153–166 (2012). doi:10.1002/dneu.20929
20. Whitfield, C.W., Ben-Shahar, Y., Brillet, C., Leoncini, I., Crauser, D., Leconte, Y., Rodriguez-Zas, S., Robinson, G.E.: Genomic dissection of behavioral maturation in the honey bee. Proc Natl Acad Sci U S A **103**(44), 16068–16075 (2006). doi:10.1073/pnas.0606909103
21. Han, B., Fang, Y., Feng, M., Hu, H., Hao, Y., Ma, C., Huo, X., Meng, L., Zhang, X., Wu, F., Li, J.: Brain membrane proteome and phosphoproteome reveal molecular basis associating with nursing and foraging behaviors of honeybee workers. Journal of proteome research **16**(10), 3646–3663 (2017)
22. Bezabih, G., Cheng, H., Han, B., Feng, M., Xue, Y., Hu, H., Li, J.: Phosphoproteome analysis reveals phosphorylation underpinnings in the brains of nurse and forager honeybees (apis mellifera). Scientific Reports **7**(1), 1973 (2017)
23. Behura, S.K., Whitfield, C.W.: Correlated expression patterns of microRNA genes with age-dependent behavioural changes in honeybee. Insect Mol Biol **19**(4), 431–439 (2010). doi:10.1111/j.1365-2583.2010.01010.x
24. Liu, F., Peng, W., Li, Z., Li, W., Li, L., Pan, J., Zhang, S., Miao, Y., Chen, S., Su, S.: Next-generation small RNA sequencing for microRNAs profiling in Apis mellifera: comparison between nurses and foragers. Insect molecular biology **21**(3), 297–303 (2012). doi:10.1111/j.1365-2583.2012.01135.x
25. Weaver, D.B., Anzola, J.M., Evans, J.D., Reid, J.G., Reese, J.T., Childs, K.L., Zdobnov, E.M., Samanta, M.P., Miller, J., Elsik, C.G.: Computational and transcriptional evidence for microRNAs in the honey bee genome. Genome biology **8**(6), 97 (2007). doi:10.1186/gb-2007-8-6-r97
26. Memczak, S., Jens, M., Elefsinioti, A., Torti, F., Krueger, J., Rybak, A., Maier, L., Mackowiak, S.D., Gregersen, L.H., Munschauer, M., Loewer, A., Ziebold, U., Landthaler, M., Kocks, C., le Noble, F., Rajewsky, N.: Circular RNAs are a large class of animal RNAs with regulatory potency. Nature **495**(7441), 333–338 (2013). doi:10.1038/nature11928
27. Shen, T., Han, M., Wei, G., Ni, T.: An intriguing RNA species–perspectives of circularized rna. Protein & cell **6**(12), 871–880 (2015). doi:10.1007/s13238-015-0202-0
28. Jeck, W.R., Sharpless, N.E.: Detecting and characterizing circular RNAs. Nature biotechnology **32**(5), 453–461 (2014). doi:10.1038/nbt.2890
29. Salzman, J., Chen, R.E., Olsen, M.N., Wang, P.L., Brown, P.O.: Cell-type specific features of circular RNA expression. PLoS genetics **9**(9), 1003777 (2013). doi:10.1371/journal.pgen.1003777
30. Ashwal-Fluss, R., Meyer, M., Pamudurti, N.R., Ivanov, A., Bartok, O., Hanan, M., Evantal, N., Memczak, S., Rajewsky, N., Kadener, S.: circRNA biogenesis competes with pre-mRNA splicing. Molecular cell **56**(1), 55–66 (2014). doi:10.1016/j.molcel.2014.08.019
31. Hansen, T.B., Jensen, T.I., Clausen, B.H., Bramsen, J.B., Finsen, B., Damgaard, C.K., Kjems, J.: Natural RNA circles function as efficient microRNA sponges. Nature **495**(7441), 384–388 (2013). doi:10.1038/nature11993
32. Gan, H., Feng, T., Wu, Y., Liu, C., Xia, Q., Cheng, T.: Identification of circular RNA in the Bombyx mori silk gland. Insect biochemistry and molecular biology **89**, 97–106 (2017). doi:10.1016/j.ibmb.2017.09.003
33. Westholm, J.O., Miura, P., Olson, S., Shenker, S., Joseph, B., Sanfilippo, P., Celniker, S.E., Graveley, B.R., Lai, E.C.: Genome-wide analysis of drosophila circular RNAs reveals their structural and sequence properties and age-dependent neural accumulation. Cell reports **9**(5), 1966–1980 (2014). doi:10.1016/j.celrep.2014.10.062
34. Kramer, M.C., Liang, D., Tatomer, D.C., Gold, B., March, Z.M., Cherry, S., Wilusz, J.E.: Combinatorial control of Drosophila circular RNA expression by intronic repeats, hnrnps, and sr proteins. Genes & development **29**(20), 2168–2182 (2015). doi:10.1101/gad.270421.115
35. Suzuki, H., Zuo, Y., Wang, J., Zhang, M.Q., Malhotra, A., Mayeda, A.: Characterization of RNase R-digested cellular RNA source that consists of lariat and circular RNAs from pre-mRNA splicing. Nucleic acids research **34**(8), 63 (2006). doi:10.1093/nar/gkl151
36. Vincent, H.A., Deutscher, M.P.: Substrate recognition and catalysis by the exoribonuclease RNase R. The Journal of biological chemistry **281**(40), 29769–29775 (2006). doi:10.1074/jbc.M606744200
37. Zeng, X., Lin, W., Guo, M., Zou, Q.: A comprehensive overview and evaluation of circular RNA detection tools. PLoS computational biology **13**(6), 1005420 (2017). doi:10.1371/journal.pcbi.1005420
38. Shen, Y., Guo, X., Wang, W.: Identification and characterization of circular RNAs in zebrafish. FEBS letters **591**(1), 213–220 (2017). doi:10.1002/1873-3468.12500
39. Elsik, C.G., Tayal, A., Diesh, C.M., Unni, D.R., Emery, M.L., Nguyen, H.N., Hagen, D.E.: Hymenoptera Genome Database: integrating genome annotations in HymenopteraMine. Nucleic Acids Research **44**(D1), 793–800 (2016). doi:10.1093/nar/gkv1208
40. Wiegmann, B.M., Trautwein, M.D., Kim, J.-W., Cassel, B.K., Bertone, M.A., Winterton, S.L., Yeates, D.K.: Single-copy nuclear genes resolve the phylogeny of the holometabolous insects. BMC biology **7**, 34 (2009). doi:10.1186/1741-7007-7-34
41. Guo, J.U., Agarwal, V., Guo, H., Bartel, D.P.: Expanded identification and characterization of mammalian circular RNAs. Genome biology **15**(7), 409 (2014). doi:10.1186/s13059-014-0409-z
42. Naeger, N.L., Van Nest, B.N., Johnson, J.N., Boyd, S.D., Southey, B.R., Rodriguez-Zas, S.L., Moore, D., Robinson, G.E.: Neurogenomic signatures of spatiotemporal memories in time-trained forager honey bees. The Journal of experimental biology **214**(Pt 6), 979–987 (2011). doi:10.1242/jeb.053421
43. Jeck, W.R., Sorrentino, J.A., Wang, K., Slevin, M.K., Burd, C.E., Liu, J., Marzluff, W.F., Sharpless, N.E.: Circular RNAs are abundant, conserved, and associated with alu repeats. RNA (New York, N.Y.) **19**(2), 141–157 (2013). doi:10.1261/rna.035667.112
44. Liang, D., Wilusz, J.E.: Short intronic repeat sequences facilitate circular RNA production. Genes & development **28**(20), 2233–2247 (2014). doi:10.1101/gad.251926.114
45. Starke, S., Jost, I., Rossbach, O., Schneider, T., Schreiner, S., Hung, L.-H., Bindereif, A.: Exon circularization requires canonical splice signals. Cell reports **10**(1), 103–111 (2015). doi:10.1016/j.celrep.2014.12.002
46. Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., Madden, T.L.: Blast+: architecture and applications. BMC

bioinformatics **10**, 421 (2009). doi:10.1186/1471-2105-10-421
47. Lorenz, R., Bernhart, S.H., Zu Siederdissen, C.H., Tafer, H., Flamm, C., Stadler, P.F., Hofacker, I.L.: ViennaRNA Package 2.0. Algorithms for Molecular Biology **6**(1), 26 (2011)
48. Shukla, S., Kavak, E., Gregory, M., Imashimizu, M., Shutinoski, B., Kashlev, M., Oberdoerffer, P., Sandberg, R., Oberdoerffer, S.: CTCF-promoted RNA polymerase II pausing links DNA methylation to splicing. Nature **479**(7371), 74–79 (2011). doi:10.1038/nature10442
49. Li-Byarlay, H., Li, Y., Stroud, H., Feng, S., Newman, T.C., Kaneda, M., Hou, K.K., Worley, K.C., Elsik, C.G., Wickline, S.A., Jacobsen, S.E., Ma, J., Robinson, G.E.: RNA interference knockdown of DNA methyl-transferase 3 affects gene alternative splicing in the honey bee. Proceedings of the National Academy of Sciences **110**(31), 12750–12755 (2013). doi:10.1073/pnas.1310735110
50. Oka, M., Rodić, N., Graddy, J., Chang, L.-J., Terada, N.: CpG sites preferentially methylated by Dnmt3a in vivo. The Journal of biological chemistry **281**(15), 9901–9908 (2006). doi:10.1074/jbc.M511100200
51. Lyko, F., Foret, S., Kucharski, R., Wolf, S., Falckenhayn, C., Maleszka, R.: The honey bee epigenomes: differential methylation of brain DNA in queens and workers. PLoS biology **8**(11), 1000506 (2010). doi:10.1371/journal.pbio.1000506
52. Becker, N., Kucharski, R., Rössler, W., Maleszka, R.: Age-dependent transcriptional and epigenomic responses to light exposure in the honey bee brain. FEBS open bio **6**(7), 622–639 (2016). doi:10.1002/2211-5463.12084
53. Herb, B.R., Wolschin, F., Hansen, K.D., Aryee, M.J., Langmead, B., Irizarry, R., Amdam, G.V., Feinberg, A.P.: Reversible switching between epigenetic states in honeybee behavioral subcastes. Nature neuroscience **15**(10), 1371–1373 (2012). doi:10.1038/nn.3218
54. Cingolani, P., Cao, X., Khetani, R.S., Chen, C.-C., Coon, M., Sammak, A., Bollig-Fischer, A., Land, S., Huang, Y., Hudson, M.E., Garfinkel, M.D., Zhong, S., Robinson, G.E., Ruden, D.M.: Intronic non-CG DNA hydroxymethylation and alternative mRNA splicing in honey bees. BMC genomics **14**, 666 (2013). doi:10.1186/1471-2164-14-666
55. Chen, H., Li, Y., Chen, K., Yao, Q., Li, G., Wang, L.: Comparative proteomic analysis of Bombyx mori hemolymph and fat body after calorie restriction. Acta biochimica Polonica **57**(4), 505–511 (2010)
56. Rybak-Wolf, A., Stottmeister, C., Glažar, P., Jens, M., Pino, N., Giusti, S., Hanan, M., Behm, M., Bartok, O., Ashwal-Fluss, R., Herzog, M., Schreyer, L., Papavasileiou, P., Ivanov, A., Öhman, M., Refojo, D., Kadener, S., Rajewsky, N.: Circular RNAs in the mammalian brain are highly abundant, conserved, and dynamically expressed. Molecular cell **58**(5), 870–885 (2015). doi:10.1016/j.molcel.2015.03.027
57. Gandini, M.A., Felix, R.: Functional interactions between voltage-gated Ca 2+ channels and Rab3-interacting molecules (RIMs): New insights into stimulus–secretion coupling. Biochimica et Biophysica Acta (BBA)-Biomembranes **1818**(3), 551–558 (2012). doi:10.1016/j.bbamem.2011.12.011
58. Garner, C.C., Kindler, S., Gundelfinger, E.D.: Molecular determinants of presynaptic active zones. Current opinion in neurobiology **10**(3), 321–327 (2000). doi:10.1016/S0959-4388(00)00093-3
59. Folkers, E., Waddell, S., Quinn, W.G.: The Drosophila radish gene encodes a protein required for anesthesia-resistant memory. Proceedings of the National Academy of Sciences of the United States of America **103**(46), 17496–17500 (2006). doi:10.1073/pnas.0608377103
60. Mery, F., Kawecki, T.J.: A cost of long-term memory in Drosophila. Science (New York, N.Y.) **308**(5725), 1148 (2005). doi:10.1126/science.1111331
61. Tully, T., Preat, T., Boynton, S.C., Del Vecchio, M.: Genetic dissection of consolidated memory in Drosophila. Cell **79**(1), 35–47 (1994)
62. Kohno, H., Suenami, S., Takeuchi, H., Sasaki, T., Kubo, T.: Production of knockout mutants by CRISPR/Cas9 in the european honeybee, Apis mellifera l. Zoological science **33**(5), 505–512 (2016). doi:10.2108/zs160043
63. Scholl, C., Kübert, N., Muenz, T.S., Rössler, W.: CaMKII knockdown affects both early and late phases of olfactory long-term memory in the honeybee. Journal of Experimental Biology **218**(23), 3788–3796 (2015). doi:10.1242/jeb.124859. http://jeb.biologists.org/content/218/23/3788.full.pdf
64. Hoffmann, S., Otto, C., Doose, G., Tanzer, A., Langenberger, D., Christ, S., Kunz, M., Holdt, L.M., Teupser, D., Hackermüller, J., Stadler, P.F.: A multi-split mapping algorithm for circular RNA, splicing, trans-splicing and fusion detection. Genome biology **15**(2), 34 (2014). doi:10.1186/gb-2014-15-2-r34
65. Li, H., Durbin, R.: Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics **25**(14), 1754–1760 (2009). doi:10.1093/bioinformatics/btp324
66. Gao, Y., Zhang, J., Zhao, F.: Circular RNA identification based on multiple seed matching. Briefings in bioinformatics (2017). doi:10.1093/bib/bbx014
67. Reim, T., Thamm, M., Rolke, D., Blenau, W., Scheiner, R.: Suitability of three common reference genes for quantitative real-time PCR in honey bees. Apidologie **44**(3), 342–350 (2013). doi:10.1007/s13592-012-0184-3
68. Zdobnov, E.M., Tegenfeldt, F., Kuznetsov, D., Waterhouse, R.M., Simão, F.A., Ioannidis, P., Seppey, M., Loetscher, A., Kriventseva, E.V.: OrthoDB v9. 1: cataloging evolutionary and functional annotations for animal, fungal, plant, archaeal, bacterial and viral orthologs. Nucleic Acids Research **45**(D1), 744–749 (2016)
69. Krueger, F., Andrews, S.R.: Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. Bioinformatics (Oxford, England) **27**(11), 1571–1572 (2011). doi:10.1093/bioinformatics/btr167
70. Langmead, B., Salzberg, S.L.: Fast gapped-read alignment with Bowtie 2. Nature methods **9**(4), 357–359 (2012). doi:10.1038/nmeth.1923
71. Kozomara, A., Griffiths-Jones, S.: miRBase: annotating high confidence microRNAs using deep sequencing data. Nucleic Acids Research **42**(D1), 68–73 (2014). doi:10.1093/nar/gkt1181
72. Lewis, B.P., Burge, C.B., Bartel, D.P.: Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. cell **120**(1), 15–20 (2005). doi:10.1016/j.cell.2004.12.035

Additional Files

Additional File 1 — Splicing
PDF showing the splice site motif of circRNAs and details on exceptions.

Additional File 2 — List of circRNAs in honeybee identified here
Excel table of all circRNAs identified here. The data is presented analogous to Table 1 but addressing additional information and details for all RNA-Seq libraries. A second sheet contains a similar list including all $3,384$ circRNAs identified by both algorithmic approaches based on JSRs in any library.

Additional File 3 — GO term enrichment
Excel table showing the results of a GO term enrichment analysis for the host genes of circRNAs based on homologous fruit fly genes performed with the PANTHER annotation platform. Terms with at least 5-fold over-representation and a false discovery rate (FDR) below 1% were considered. From these, we limited the interpretation with a relevant p-value threshold of 10^{-4} which is marked in the table.

Additional File 4 — List of potential miRNA targets including conservations
Excel table of all miRNA target sites found on circRNA sequences identified here. It contains detailed data on the target circRNA, the potential position of interaction and its conservation in *Apis*, eusocial insects, *Drosophila* and *Bombyx*.

Additional File 5 — List of primers and probes for PCR and TaqMan assay
Excel table with a list of all PCR primes used in this study. A second sheet lists the TaqMan probes.

Additional File 6 — Verification of circRNAs via PCR
Results of the circularity validation through PCR.

Additional File 7 — RNase R enrichment control
Experimental control of circRNA enrichment over linear products.

Bibliography

[1] Crick FH (1958): **On protein synthesis**. *Symposia of the Society for Experimental Biology*. 12: 138–63. Available: https://www.ncbi.nlm.nih.gov/pubmed/13580867?dopt=Abstract

[2] Crick F (1970): **Central Dogma of Molecular Biology**. *Nature*. Nature Publishing Group; 227: 561. doi:10.1038/227561a0

[3] Temin HM, Mizutani S (1970): **Viral RNA-dependent DNA Polymerase: RNA-dependent DNA Polymerase in Virions of Rous Sarcoma Virus**. *Nature*. Nature Publishing Group; 226: 1211. doi:10.1038/2261211a0

[4] Eddy SR (2001): **Non–coding RNA genes and the modern RNA world**. *Nature Reviews Genetics*. Nature Publishing Group; 2: 919. doi:10.1038/35103511

[5] Garst AD, Edwards AL, Batey RT (2011): **Riboswitches: Structures and mechanisms**. *Cold Spring Harbor perspectives in biology*. Cold Spring Harbor Lab; 3: a003533. doi:10.1101/cshperspect.a003533

[6] Fire A, Xu S, Montgomery MK, Kostas SA, Driver SE, Mello CC (1998): **Potent and specific genetic interference by double-stranded RNA in *Caenorhabditis elegans***. *Nature*. Nature Publishing Group; 391: 806. doi:10.1038/35888

[7] Barrick JE, Sudarsan N, Weinberg Z, Ruzzo WL, Breaker RR (2005): **6S RNA is a widespread regulator of eubacterial RNA polymerase that resembles an open promoter**. *RNA*. Cold Spring Harbor Lab; 11: 774–784. doi:10.1261/rna.7286705

[8] Guerrier-Takada C, Gardiner K, Marsh T, Pace N, Altman S (1983): **The RNA moiety of ribonuclease P is the catalytic subunit of the enzyme**. *Cell*. Cell Press; 35: 849–857. doi:10.1016/0092-8674(83)90117-4

[9] Hambraeus G, Wachenfeldt C von, Hederstedt L (2003): **Genome-wide survey of mRNA half-lives in *Bacillus subtilis* identifies extremely stable mRNAs**. *Molecular Genetics and Genomics*. Springer-Verlag; 269: 706–714. doi:10.1007/s00438-003-0883-6

[10] Blattner FR, Plunkett G, Bloch CA, Perna NT, Burland V, Riley M, Collado-Vides J, Glasner JD, Rode CK, Mayhew GF, Gregor J, Davis NW, Kirkpatrick HA, Goeden MA, Rose DJ, Mau B, Shao Y (1997): **The Complete Genome Sequence of *Escherichia coli* K-12**. *Science*. American Association for the Advancement of Science; 277: 1453–1462. doi:10.1126/science.277.5331.1453

[11] Pearson H (2006): **Genetics: What is a gene?** *Nature*. Nature Publishing Group; 441: 398. doi:10.1038/441398a

[12] Pribnow D (1975): **Nucleotide sequence of an RNA polymerase binding site at an early T7 promoter**. *Proceedings of the National Academy of Sciences*. National Academy of Sciences; 72: 784–788. doi:10.1073/pnas.72.3.784

[13] Schaller H, Gray C, Herrmann K (1975): **Nucleotide sequence of an RNA polymerase binding site from the DNA of bacteriophage fd**. *Proceedings of the National Academy of Sciences*. National Academy of Sciences; 72: 737–741. doi:10.1073/pnas.72.2.737

[14] Feklistov A, Darst SA (2011): **Structural Basis for Promoter -10 Element Recognition by the Bacterial RNA Polymerase σ Subunit**. *Cell*. Cell Press; 147: 1257–1269. doi:10.1016/j.cell.2011.10.041

[15] Gruber TM, Gross CA (2003): **Multiple Sigma Subunits and the Partitioning of Bacterial Transcription Space**. *Annual Review of Microbiology*. Annual Reviews; 57: 441–466. doi:10.1146/annurev.micro.57.030502.090913

[16] Boudvillain M, Figueroa-Bossi N, Bossi L (2013): **Terminator still moving forward: expanding roles for Rho factor**. *Current Opinion in Microbiology*. Elsevier Current Trends; 16: 118–124. doi:10.1016/j.mib.2012.12.003

[17] Kingsford CL, Ayanbule K, Salzberg SL (2007): **Rapid, accurate, computational discovery of rho-independent transcription terminators illuminates their relationship to dna uptake**. *Genome biology*. BioMed Central; 8: R22. doi:10.1186/gb-2007-8-2-r22

[18] Mingorance J, Tamames J, Vicente M (2004): **Genomic channeling in bacterial cell division**. *Journal of molecular recognition*. Wiley Online Library; 17: 481–487. doi:10.1002/jmr.718

[19] Tamames J, González-Moreno M, Mingorance J, Valencia A, Vicente M (2001): **Bringing gene order into bacterial shape**. *Trends in Genetics*. Elsevier; 17: 124–126. doi:10.1016/S0168-9525(00)02212-5

[20] Vicente M, Gomez M, Ayala J (1998): **Regulation of transcription of cell division genes in the escherichia coli dcw cluster**. *Cellular and molecular life sciences*. Springer; 54: 317–324. doi:10.1007/s000180050158

[21] Mingorance J, Tamames J (2004): **The bacterial dcw gene cluster: an island in the genome?** *Molecules in Time and Space*. Springer, Boston, MA; 249–271. doi:10.1007/0-306-48579-6_13

[22] Duez C, Thamm I, Sapunaric F, Coyette J, Ghuysen JM (1998): **The Division and Cell Wall Gene Cluster of *Enterococcus hirae* S185**. *DNA Sequence*. Taylor & Francis; 9: 149–161. doi:10.3109/10425179809072190

[23] Berghoff BA, Glaeser J, Sharma CM, Vogel J, Klug G (2009): **Photooxidative stress-induced and abundant small RNAs in *Rhodobacter sphaeroides***. *Molecular microbiology*. Wiley Online Library; 74: 1497–1512. doi:10.1111/j.1365-2958.2009.06949.x

[24] Fuente A de la, Palacios P, Vicente M (2001): **Transcription of the *Escherichia coli* dcw cluster: Evidence for distal upstream transcripts being involved in the expression of the downstream ftsZ gene**. *Biochimie*. Elsevier; 83: 109–115. doi:10.1016/S0300-9084(00)01212-8

[25] Naylor GW, Addlesee HA, Gibson LCD, Hunter CN (1999): **The photosynthesis gene cluster of *Rhodobacter sphaeroides***. *Photosynthesis Research*. Kluwer Academic Publishers; 62: 121–139. doi:10.1023/A:1006350405674

[26] Brownlee GG (1971): **Sequence of 6S RNA of *E. coli***. *Nature New Biology*. Nature Publishing Group UK; 229: 147–149. doi:10.1038/newbio229147a0

[27] Ando Y, Asari S, Suzuma S, Yamane K, Nakamura K (2002): **Expression of a small RNA, BS203 RNA, from the yocI–yocJ intergenic region of *Bacillus subtilis* genome**. *FEMS Microbiology Letters*. Oxford University Press; 207: 29–33. doi:10.1111/j.1574-6968.2002.tb11023.x

[28] Burenina OY, Hoch PG, Damm K, Salas M, Zatsepin TS, Lechner M, Oretskaya TS, Kubareva EA, Hartmann RK (2014): **Mechanistic comparison of *Bacillus subtilis* 6S-1 and 6S-2 RNAs – commonalities and differences**. *RNA*. Cold Spring Harbor Lab; doi:10.1261/rna.042077.113

[29] Beckmann BM, Burenina OY, Hoch PG, Kubareva EA, Sharma CM, Hartmann RK (2011): **In vivo and in vitro analysis of 6S RNA-templated short transcripts in *Bacillus subtilis***. *RNA Biology*. Taylor & Francis; 8: 839–849. doi:10.4161/rna.8.5.16151

[30] Beckmann BM, Hoch PG, Marz M, Willkomm DK, Salas M, Hartmann RK (2012): **A pRNA-induced structural rearrangement triggers 6S-1 RNA release from RNA polymerase in *Bacillus subtilis***. *EMBO Journal*. EMBO Press; 31: 1727–1738. doi:10.1038/emboj.2012.23

[31] Laalami S, Zig L, Putzer H (2014): **Initiation of mRNA decay in bacteria**. *Cellular and Molecular Life Sciences*. Springer Basel; 71: 1799–1828. doi:10.1007/s00018-013-1472-4

[32] Condon C (2003): **RNA Processing and Degradation in *Bacillus subtilis***. *Microbiol Mol Biol Rev*. American Society for Microbiology; 67: 157–174. doi:10.1128/MMBR.67.2.157-174.2003

[33] Craven MG, Henner DJ, Alessi D, Schauer AT, Ost KA, Deutscher MP, Friedman DI (1992): **Identification of the *rph* (RNase PH) gene of *Bacillus subtilis*: evidence for suppression of cold-sensitive mutations in *Escherichia coli*.** *Journal of Bacteriology*. American Society for Microbiology Journals; 174: 4727–4735. doi:10.1128/jb.174.14.4727-4735.1992

[34] Luttinger A, Hahn J, Dubnau D (1996): **Polynucleotide phosphorylase is necessary for competence development in *Bacillus subtilis***. *Molecular Microbiology*. Wiley/Blackwell (10.1111); 19: 343–356. doi:10.1046/j.1365-2958.1996.380907.x

[35] Mitra S, Hue K, Bechhofer DH (1996): **In vitro processing activity of *Bacillus subtilis* polynucleotide phosphorylase**. *Molecular Microbiology*. Wiley/Blackwell (10.1111); 19: 329–342. doi:10.1046/j.1365-2958.1996.378906.x

[36] Mathy N, Bénard L, Pellegrini O, Daou R, Wen T, Condon C (2007): **5′-to-3′ Exoribonuclease Activity in Bacteria: Role of RNase J1 in rRNA Maturation and 5′ Stability of mRNA**. *Cell*. Cell Press; 129: 681–692. doi:10.1016/j.cell.2007.02.051

[37] Even S, Pellegrini O, Zig L, Labas V, Vinh J, Bréchemmier-Baey D, Putzer H (2005): **Ribonucleases J1 and J2: two novel endoribonucleases in *B.subtilis* with functional homology to *E.coli* RNase E**. *Nucleic Acids Research*. Oxford University Press; 33: 2141–2152. doi:10.1093/nar/gki505

[38] Durand S, Gilet L, Bessières P, Nicolas P, Condon C (2012): **Three Essential Ribonucleases – RNase Y, J1, and III – Control the Abundance of a Majority of *Bacillus subtilis* mRNAs**. *PLOS Genetics*. Public Library of Science; 8: e1002520. doi:10.1371/journal.pgen.1002520

[39] Sierra-Gallay IL de la, Zig L, Jamalli A, Putzer H (2008): **Structural insights into the dual activity of RNase J**. *Nature Structural & Molecular Biology*. Nature Publishing Group; 15: 206. doi:10.1038/nsmb.1376

[40] Laalami S, Bessières P, Rocca A, Zig L, Nicolas P, Putzer H (2013): ***Bacillus subtilis* RNase Y Activity In Vivo Analysed by Tiling Microarrays**. *PLOS ONE*. Public Library of Science; 8: e54062. doi:10.1371/journal.pone.0054062

[41] Nikolaev N, Silengo L, Schlessinger D (1973): **A Role for Ribonuclease III in Processing of Ribosomal Ribonucleic Acid and Messenger Ribonucleic Acid Precursors in *Escherichia coli***. *Journal of Biological Chemistry*. American Society for Biochemistry; Molecular Biology; 248: 7967–7969. Available: http://www.jbc.org/content/248/22/7967.short

[42] Cheng Z-F, Zuo Y, Li Z, Rudd KE, Deutscher MP (1998): **The *vacB* Gene Required for Virulence in *Shigella flexneri* and *Escherichia coli* Encodes the Exoribonuclease RNase R**. *Journal of Biological Chemistry*. American Society for Biochemistry; Molecular Biology; 273: 14077–14080. doi:10.1074/jbc.273.23.14077

[43] Oussenko IA, Sanchez R, Bechhofer DH (2002): ***Bacillus subtilis* YhaM, a Member of a New Family of 3′-to-5′ Exonucleases in Gram-Positive Bacteria**. *Journal of Bacteriology*. American Society for Microbiology Journals; 184: 6250–6259. doi:10.1128/JB.184.22.6250-6259.2002

[44] Sprinzl M, Horn C, Brown M, Ioudovitch A, Steinberg S (1998): **Compilation of tRNA sequences and sequences of tRNA genes**. *Nucleic Acids Research*. Oxford University Press; 26: 148–153. doi:10.1093/nar/26.1.148

[45] Lai LB, Vioque A, Kirsebom LA, Gopalan V (2010): **Unexpected diversity of RNase P, an ancient tRNA processing enzyme: Challenges and prospects**. *FEBS Letters*. No longerElsevier; 584: 287–296. doi:10.1016/j.febslet.2009.11.048

[46] Hartmann RK, Gößringer M, Späth B, Fischer S, Marchfelder A (2009): **Chapter 8 The Making of tRNAs and More – RNase P and tRNase Z**. *Progress in Molecular Biology and Translational Science*. Academic Press; 85: 319–368. doi:10.1016/S0079-6603(08)00808-8

[47] Jarrous N, Gopalan V (2010): **Archaeal/Eukaryal RNase P: subunits, functions and RNA diversification**. *Nucleic Acids Research*. Oxford University Press; 38: 7885–7894. doi:10.1093/nar/gkq701

[48] Marvin MC, Engelke DR (2009): **RNase P: increased versatility through protein complexity?** *RNA Biology*. Taylor & Francis; 6: 40–42. doi:10.4161/rna.6.1.7566

[49] Marquez SM, Harris JK, Kelley ST, Brown JW, Dawson SC, Roberts EC, Pace NR (2005): **Structural implications of novel diversity in eucaryal RNase P RNA**. *RNA*. Cold Spring Harbor Lab; 11: 739–751. doi:10.1261/rna.7211705

[50] Piccinelli P, Rosenblad MA, Samuelsson T (2005): **Identification and analysis of ribonuclease P and MRP RNA in a broad range of eukaryotes**. *Nucleic Acids Research*. Oxford University Press; 33: 4485–4495. doi:10.1093/nar/gki756

[51] Marvin MC, Engelke DR (2009): **Broadening the mission of an RNA enzyme**. *Journal of Cellular Biochemistry*. Wiley-Blackwell; 108: 1244–1251. doi:10.1002/jcb.22367

[52] Kikovska E, Svärd SG, Kirsebom LA (2007): **Eukaryotic RNase P RNA mediates cleavage in the absence of protein**. *Proceedings of the National Academy of Sciences*. National Academy of Sciences; 104: 2062–2067. doi:10.1073/pnas.0607326104

[53] Willkomm DK, Hartmann RK (2007): **An important piece of the RNase P jigsaw solved**. *Trends in Biochemical Sciences*. Elsevier Current Trends; 32: 247–250. doi:10.1016/j.tibs.2007.04.005

[54] Chang D, Clayton D (1987): **A mammalian mitochondrial RNA processing activity contains nucleus-encoded RNA**. *Science*. American Association for the Advancement of Science; 235: 1178–1184. doi:10.1126/science.2434997

[55] Holzmann J, Frank P, Löffler E, Bennett KL, Gerner C, Rossmanith W (2008): **RNase P without RNA: Identification and Functional Reconstitution of the Human Mitochondrial tRNA Processing Enzyme**. *Cell*. Cell Press; 135: 462–474. doi:10.1016/j.cell.2008.09.013

[56] Taschner A, Weber C, Buzet A, Hartmann RK, Hartig A, Rossmanith W (2012): **Nuclear RNase P of *Trypanosoma brucei*: A Single Protein in Place of the Multicomponent RNA-Protein Complex**. *Cell Reports*. Cell Press; 2: 19–25. doi:10.1016/j.celrep.2012.05.021

[57] Bicknell AA, Cenik C, Chua HN, Roth FP, Moore MJ (2012): **Introns in UTRs: Why we should stop ignoring them**. *BioEssays*. Wiley-Blackwell; 34: 1025–1034. doi:10.1002/bies.201200073

[58] Gagniuc P, Ionescu-Tirgoviste C (2012): **Eukaryotic genomes may exhibit up to 10 generic classes of gene promoters**. *BMC Genomics*. BioMed Central; 13: 512. doi:10.1186/1471-2164-13-512

[59] Jurica MS, Moore MJ (2003): **Pre-mRNA Splicing: Awash in a Sea of Proteins**. *Molecular Cell*. Cell Press; 12: 5–14. doi:10.1016/S1097-2765(03)00270-3

[60] Blencowe BJ (2006): **Alternative Splicing: New Insights from Global Analyses**. *Cell*. Cell Press; 126: 37–47. doi:10.1016/j.cell.2006.06.023

[61] Memczak S, Jens M, Elefsinioti A, Torti F, Krueger J, Rybak A, Maier L, Mackowiak SD, Gregersen LH, Munschauer M, Loewer A, Ziebold U, Landthaler M, Kocks C, Noble F le, Rajewsky N (2013): **Circular RNAs are a large class of animal RNAs with regulatory potency**. *Nature*. 495: 333–338. doi:10.1038/nature11928

[62] Sanger HL, Klotz G, Riesner D, Gross HJ, Kleinschmidt AK (1976): **Viroids are single-stranded covalently closed circular RNA molecules existing as highly base-paired rod-like structures**. *Proceedings of the National Academy of Sciences*. National Acad Sciences; 73: 3852–3856. doi:10.1073/pnas.73.11.3852

[63] Hsu M-T, Coca-Prados M (1979): **Electron microscopic evidence for the circular form of RNA in the cytoplasm of eukaryotic cells**. *Nature*. Springer; 280: 339–340. doi:10.1038/280339a0

[64] Nigro JM, Cho KR, Fearon ER, Kern SE, Ruppert JM, Oliner JD, Kinzler KW, Vogelstein B (1991): **Scrambled exons**. *Cell*. Elsevier; 64: 607–613. doi:10.1016/0092-8674(91)90244-S

[65] Capel B, Swain A, Nicolis S, Hacker A, Walter M, Koopman P, Goodfellow P, Lovell-Badge R (1993): **Circular transcripts of the testis-determining gene *sry* in adult mouse testis**. *Cell*. Elsevier; 73: 1019–1030. doi:10.1016/0092-8674(93)90279-Y

[66] Cocquerelle C, Mascrez B, Hétuin D, Bailleul B (1993): **Mis-splicing yields circular RNA molecules**. *FASEB journal : official publication of the Federation of American Societies for Experimental Biology*. 7: 155–160.

[67] Suzuki H, Zuo Y, Wang J, Zhang MQ, Malhotra A, Mayeda A (2006): **Characterization of RNase R-digested cellular RNA source that consists of lariat and circular RNAs from pre-mRNA splicing**. *Nucleic acids research*. Oxford University Press; 34: e63–e63. doi:10.1093/nar/gkl151

[68] Westholm JO, Miura P, Olson S, Shenker S, Joseph B, Sanfilippo P, Celniker SE, Graveley BR, Lai EC (2014): **Genome-wide analysis of *Drosophila* circular RNAs reveals their structural and sequence properties and age-dependent neural accumulation**. *Cell reports*. 9: 1966–1980. doi:10.1016/j.celrep.2014.10.062

[69] Yang Y, Fan X, Mao M, Song X, Wu P, Zhang Y, Jin Y, Yang Y, Chen L-L, Wang Y, others (2017): **Extensive translation of circular RNAs driven by N 6-methyladenosine**. *Cell research*. Nature Publishing Group; 27: 626. doi:10.1038/cr.2017.31

[70] Liang D, Tatomer DC, Luo Z, Wu H, Yang L, Chen L-L, Cherry S, Wilusz JE (2017): **The output of protein-coding genes shifts to circular RNAs when the pre-mRNA processing machinery is limiting**. *Molecular cell*. Elsevier; 68: 940–954. doi:10.1016/j.molcel.2017.10.034

[71] You X, Vlatkovic I, Babic A, Will T, Epstein I, Tushev G, Akbalik G, Wang M, Glock C, Quedenau C, others (2015): **Neural circular RNAs are derived from synaptic genes and regulated by development and plasticity**. *Nature neuroscience*. Nature Research; 18: 603–610. doi:10.1038/nn.3975

[72] Holdt LM, Stahringer A, Sass K, Pichler G, Kulak NA, Wilfert W, Kohlmaier A, Herbst A, Northoff BH, Nicolaou A, Gäbel G, Beutner F, Scholz M, Thiery J, Musunuru K, Krohn K, Mann M, Teupser D (2016): **Circular non-coding RNA ANRIL modulates ribosomal RNA maturation and atherosclerosis in humans**. *Nature communications*. Nature Publishing Group; 7: 12429. doi:10.1038/ncomms12429

[73] Ashwal-Fluss R, Meyer M, Pamudurti NR, Ivanov A, Bartok O, Hanan M, Evantal N, Memczak S, Rajewsky N, Kadener S (2014): **CircRNA biogenesis competes with pre-mRNA splicing**. *Molecular cell*. 56: 55–66. doi:10.1016/j.molcel.2014.08.019

[74] Gao Y, Wang J, Zheng Y, Zhang J, Chen S, Zhao F (2016): **Comprehensive identification of internal structure and alternative splicing events in circular RNAs**. *Nature communications*. Nature Publishing Group; 7. doi:10.1038/ncomms12060

[75] Fica SM, Tuttle N, Novak T, Li N-S, Lu J, Koodathingal P, Dai Q, Staley JP, Piccirilli JA (2013): **RNA catalyses nuclear pre-mRNA splicing**. *Nature*. Nature Publishing Group; 503: 229. doi:10.1038/nature12734

[76] Szabo L, Salzman J (2016): **Detecting circular RNAs: Bioinformatic and experimental challenges**. *Nature reviews Genetics*. Nature Research; 17: 679–692. doi:10.1038/nrg.2016.114

[77] Li Z, Huang C, Bao C, Chen L, Lin M, Wang X, Zhong G, Yu B, Hu W, Dai L, others (2015): **Exon-intron circular RNAs regulate transcription in the nucleus**. *Nature structural & molecular biology*. Nature Research; 22: 256–264. doi:10.1038/nsmb.2959

[78] Ivanov A, Memczak S, Wyler E, Torti F, Porath HT, Orejuela MR, Piechotta M, Levanon EY, Landthaler M, Dieterich C, others (2015): **Analysis of intron sequences reveals hallmarks of circular RNA biogenesis in animals**. *Cell reports*. Elsevier; 10: 170–177. doi:10.1016/j.celrep.2014.12.019

[79] Starke S, Jost I, Rossbach O, Schneider T, Schreiner S, Hung L-H, Bindereif A (2015): **Exon circularization requires canonical splice signals**. *Cell reports*. Elsevier; 10: 103–111. doi:10.1016/j.celrep.2014.12.002

[80] Wilusz JE (2015): **Repetitive elements regulate circular RNA biogenesis**. *Mobile genetic elements*. Taylor & Francis; 5: 39–45. doi:10.1080/2159256X.2015.1045682

[81] Kramer MC, Liang D, Tatomer DC, Gold B, March ZM, Cherry S, Wilusz JE (2015): **Combinatorial control of *Drosophila* circular RNA expression by intronic repeats, hnRNPs, and sr proteins**. *Genes & development*. 29: 2168–2182. doi:10.1101/gad.270421.115

[82] Dong R, Ma X-K, Chen L-L, Yang L (2017): **Increased complexity of circRNA expression during species evolution**. *RNA Biology*. Taylor & Francis; 14: 1064–1074. doi:10.1080/15476286.2016.1269999

[83] Shen Y, Guo X, Wang W (2017): **Identification and characterization of circular RNAs in zebrafish**. *FEBS letters*. Wiley Online Library; 591: 213–220. doi:10.1002/1873-3468.12500

[84] Pamudurti NR, Bartok O, Jens M, Ashwal-Fluss R, Stottmeister C, Ruhe L, Hanan M, Wyler E, Perez-Hernandez D, Ramberger E, others (2017): **Translation of circRNAs**. *Molecular Cell*. Elsevier; 66: 9–21. doi:10.1016/j.molcel.2017.02.021

[85] Hansen TB, Jensen TI, Clausen BH, Bramsen JB, Finsen B, Damgaard CK, Kjems J (2013): **Natural RNA circles function as efficient microRNA sponges**. *Nature*. 495: 384–388. doi:10.1038/nature11993

[86] Guo JU, Agarwal V, Guo H, Bartel DP (2014): **Expanded identification and characterization of mammalian circular RNAs**. *Genome biology*. BioMed Central; 15: 409. doi:10.1186/s13059-014-0409-z

[87] Gan H, Feng T, Wu Y, Liu C, Xia Q, Cheng T (2017): **Identification of circular RNA in the *Bombyx mori* silk gland**. *Insect biochemistry and molecular biology*. Elsevier; 89: 97–106. doi:10.1016/j.ibmb.2017.09.003

[88] Winston ML (1991.): **The biology of the honey bee**. Harvard University Press;

[89] Seeley T (1995.): **The wisdom of the hive**. Cambridge Mass, London: Harvard University Press;

[90] Ben-Shahar Y, Thompson CK, Hartz SM, Smith BH, Robinson GE (2000): **Differences in performance on a reversal learning test and division of labor in honey bee colonies**. *Animal Cognition*. Springer-Verlag; 3: 119–125. doi:10.1007/s100710000068

[91] Robinson GE, Page RE, Strambi C, Strambi A (1992): **Colony Integration in Honey Bees: Mechanisms of Behavioral Reversion**. *Ethology*. Wiley/Blackwell (10.1111); 90: 336–348. doi:10.1111/j.1439-0310.1992.tb00844.x

[92] Herb BR, Wolschin F, Hansen KD, Aryee MJ, Langmead B, Irizarry R, Amdam GV, Feinberg AP (2012): **Reversible switching between epigenetic states in honeybee behavioral subcastes**. *Nature neuroscience*. Nature Publishing Group; 15: 1371. doi:10.1038/nn.3218

[93] Welsh L, Maleszka R, Foret S (2017): **Detecting rare asymmetrically methylated cytosines and decoding methylation patterns in the honeybee genome**. *Royal Society open science*. The Royal Society; 4: 170248. doi:10.1098/rsos.170248

[94] Cohen NM, Kenigsberg E, Tanay A (2011): **Primate CpG Islands Are Maintained by Heterogeneous Evolutionary Regimes Involving Minimal Selection**. *Cell*. Cell Press; 145: 773–786. doi:10.1016/j.cell.2011.04.024

[95] Bird AP (1980): **DNA methylation and the frequency of CpG in animal DNA**. *Nucleic Acids Research*. Oxford University Press; 8: 1499–1504. doi:10.1093/nar/8.7.1499

[96] Saxonov S, Berg P, Brutlag DL (2006): **A genome-wide analysis of CpG dinucleotides in the human genome distinguishes two distinct classes of promoters**. *Proceedings of the National Academy of Sciences*. National Academy of Sciences; 103: 1412–1417. doi:10.1073/pnas.0510310103

[97] Wu SC, Zhang Y (2010): **Active DNA demethylation: many roads lead to Rome**. *Nature reviews Molecular cell biology*. Nature Publishing Group; 11: 607–620. doi:10.1038/nrm2950

[98] Ito S, D'Alessio AC, Taranova OV, Hong K, Sowers LC, Zhang Y (2010): **Role of Tet proteins in 5mC to 5hmC conversion, ES-cell self-renewal and inner cell mass specification**. *Nature*. Nature Publishing Group; 466: 1129–1133. doi:10.1038/nature09303

[99] Deaton AM, Bird A (2011): **CpG islands and the regulation of transcription**. *Genes & Development*. Cold Spring Harbor Lab; 25: 1010–1022. doi:10.1101/gad.2037511

[100] Provataris P, Meusemann K, Niehuis O, Grath S, Misof B (2018): **Signatures of dna methylation across insects suggest reduced dna methylation levels in holometabola**. *Genome biology and evolution*. Oxford University Press; 10: 1185–1197. doi:10.1093/gbe/evy066

[101] Lyko F, Foret S, Kucharski R, Wolf S, Falckenhayn C, Maleszka R (2010): **The honey bee epigenomes: Differential methylation of brain DNA in queens and workers.** *PLoS biology*. 8: e1000506. doi:10.1371/journal.pbio.1000506

[102] Kucharski R, Maleszka J, Foret S, Maleszka R (2008): **Nutritional control of reproductive status in honeybees via dna methylation**. *Science*. American Association for the Advancement of Science; 319: 1827–1830. doi:10.1126/science.1153069

[103] Li-Byarlay H, Li Y, Stroud H, Feng S, Newman TC, Kaneda M, Hou KK, Worley KC, Elsik CG, Wickline SA, others (2013): **RNA interference knockdown of dna methyl-transferase 3 affects gene alternative splicing in the honey bee**. *Proceedings of the National Academy of Sciences*. National Acad Sciences; 110: 12750–12755. doi:10.1073/pnas.1310735110

[104] Bonasio R, Li Q, Lian J, Mutti NS, Jin L, Zhao H, Zhang P, Wen P, Xiang H, Ding Y, others (2012): **Genome-wide and caste-specific dna methylomes of the ants *Camponotus floridanus* and *Harpegnathos saltator***. *Current Biology*. Elsevier; 22: 1755–1764. doi:10.1016/j.cub.2012.07.042

[105] Sanger F, Coulson AR (1975): **A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase**. *Journal of Molecular Biology*. Academic Press; 94: 441–448. doi:10.1016/0022-2836(75)90213-2

[106] Sanger F, Nicklen S, Coulson AR (1977): **DNA sequencing with chain-terminating inhibitors**. *Proceedings of the National Academy of Sciences of the United States of America*. National Academy of Sciences; 74: 5463–5467. Available: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC431765

[107] Sanger F, Air GM, Barrell BG, Brown NL, Coulson AR, Fiddes JC, Hutchison Iii CA, Slocombe PM, Smith M (1977): **Nucleotide sequence of bacteriophage φX174 DNA**. *Nature*. Nature Publishing Group; 265: 687. doi:10.1038/265687a0

[108] Beck S, Pohl FM (1984): **DNA sequencing with direct blotting electrophoresis.** *EMBO Journal*. European Molecular Biology Organization; 3: 2905–2909. Available: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC557787

[109] Craig Chinault A, Carbon J (1979): **Overlap hybridization screening: Isolation and characterization of overlapping DNA fragments surrounding the leu2 gene on yeast chromosome III**. *Gene*. Elsevier; 5: 111–126. doi:10.1016/0378-1119(79)90097-0

[110] Anderson S (1981): **Shotgun DNA sequencing using cloned DNase I-generated fragments**. *Nucleic Acids Research*. Oxford University Press; 9: 3015–3027. doi:10.1093/nar/9.13.3015

[111] Staden R (1979): **A strategy of DNA sequencing employing computer programs**. *Nucleic Acids Research*. Oxford University Press; 6: 2601–2610. doi:10.1093/nar/6.7.2601

[112] Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, Berka J, Braverman MS, Chen Y-J, Chen Z, Dewell SB, Du L, Fierro JM, Gomes XV, Godwin BC, He W, Helgesen S, Ho CH, Irzyk GP, Jando SC, Alenquer MLI, Jarvie TP, Jirage KB, Kim J-B, Knight JR, Lanza JR, Leamon JH, Lefkowitz SM, Lei M, Li J, Lohman KL, Lu H, Makhijani VB, McDade KE, McKenna MP, Myers EW, Nickerson E, Nobile JR, Plant R, Puc BP, Ronan MT, Roth GT, Sarkis GJ, Simons JF, Simpson JW, Srinivasan M, Tartaro KR, Tomasz A, Vogt KA, Volkmer GA, Wang SH, Wang Y, Weiner MP, Yu P, Begley RF, Rothberg JM (2005): **Genome sequencing in microfabricated high-density picolitre reactors**. *Nature*. Nature Publishing Group; 437: 376. doi:10.1038/nature03959

[113] Metzker ML (2009): **Sequencing technologies – the next generation**. *Nature Reviews Genetics*. Nature Publishing Group; 11: 31. doi:10.1038/nrg2626

[114] Smith LM, Sanders JZ, Kaiser RJ, Hughes P, Dodd C, Connell CR, Heiner C, Kent SBH, Hood LE (1986): **Fluorescence detection in automated DNA sequence analysis**. *Nature*. Nature Publishing Group; 321: 674. doi:10.1038/321674a0

[115] Ewing B, Hillier L, Wendl MC, Green P (1998): **Base-calling of automated sequencer traces using phred. I. Accuracy assessment**. *Genome research*. 8: 175–185. Available: https://www.ncbi.nlm.nih.gov/pubmed/9521921

[116] Adams MD, Kelley JM, Gocayne JD, Dubnick M, Polymeropoulos MH, Xiao H, Merril CR, Wu A, Olde B, Moreno RF (1991): **Complementary DNA sequencing: expressed sequence tags and human genome project**. *Science (New York, NY)*. HighWire - PDF; 252: 1651–1656. Available: https://www.ncbi.nlm.nih.gov/pubmed/2047873

[117] Kukurba KR, Montgomery SB (2015): **RNA Sequencing and Analysis**. *Cold Spring Harbor Protocols*. Cold Spring Harbor Laboratory Press; 2015: pdb.top084970. doi:10.1101/pdb.top084970

[118] Edwards A, Caskey CT (1991): **Closure strategies for random DNA sequencing**. *Methods*. Academic Press; 3: 41–47. doi:10.1016/S1046-2023(05)80162-8

[119] Wang Z, Gerstein M, Snyder M (2009): **RNA-Seq: a revolutionary tool for transcriptomics**. *Nature Reviews Genetics*. Nature Publishing Group; 10: 57. doi:10.1038/nrg2484

[120] O'Neil D, Glowatz H, Schlumpberger M (2013): **Ribosomal RNA Depletion for Efficient Use of RNA-Seq Capacity**. *Current Protocols in Molecular Biology*. Wiley-Blackwell; 103: 4.19.1–4.19.8. doi:10.1002/0471142727.mb0419s103

[121] Sharp SJ, Schaack J, Cooley L, Burke DJ, Soil D (1985): **Structure and Transcription of Eukaryotic tRNA Gene**. *Critical Reviews in Biochemistry*. Taylor & Francis; 19: 107–144. doi:10.3109/10409238509082541

[122] Vogel J, Wagner EGH (2007): **Target identification of small noncoding RNAs in bacteria**. *Current Opinion in Microbiology*. Elsevier Current Trends; 10: 262–270. doi:10.1016/j.mib.2007.06.001

[123] Lewis BP, Burge CB, Bartel DP (2005): **Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets**. *Cell*. Elsevier; 120: 15–20. doi:https://doi.org/10.1016/j.cell.2004.12.035

[124] Knierim E, Lucke B, Schwarz JM, Schuelke M, Seelow D (2011): **Systematic Comparison of Three Methods for Fragmentation of Long-Range PCR Products for Next Generation Sequencing**. *PLOS ONE*. Public Library of Science; 6: e28240. doi:10.1371/journal.pone.0028240

[125] Roach JC, Boysen C, Wang K, Hood L (1995): **Pairwise end sequencing: a unified approach to genomic mapping and sequencing**. *Genomics*. Academic Press; 26: 345–353. doi:10.1016/0888-7543(95)80219-C

[126] Sharma CM, Hoffmann S, Darfeuille F, Reignier J, Findeiß S, Sittka A, Chabas S, Reiche K, Hackermüller J, Reinhardt R, Stadler PF, Vogel J (2010): **The primary transcriptome of the major human pathogen *Helicobacter pylori***. *Nature*. Nature Publishing Group; 464: 250. doi:10.1038/nature08756

[127] Shinshi H, Miwa M, Kato K, Noguchi M, Matsushima T, Sugimura T (1976): **A novel phosphodiesterase from cultured tobacco cells**. *Biochemistry*. American Chemical Society; 15: 2185–2190. doi:10.1021/bi00655a024

[128] Babski J, Haas KA, Näther-Schindler D, Pfeiffer F, Förstner KU, Hammelmann M, Hilker R, Becker A, Sharma CM, Marchfelder A, Soppa J (2016): **Genome-wide identification of transcriptional start sites in the haloarchaeon *Haloferax volcanii* based on differential RNA-Seq (dRNA-Seq)**. *BMC Genomics*. BioMed Central; 17: 629. doi:10.1186/s12864-016-2920-y

[129] Weber L, Thoelken C, Volk M, Remes B, Lechner M, Klug G (2016): **The Conserved Dcw Gene Cluster of *R. sphaeroides* Is Preceded by an Uncommonly Extended 5' Leader Featuring the sRNA UpsM**. *PLoS One*. Public Library of Science; 11: e0165694. doi:10.1371/journal.pone.0165694

[130] Weber M, Davies JJ, Wittig D, Oakeley EJ, Haase M, Lam WL, Schübeler D (2005): **Chromosome-wide and promoter-specific analyses identify sites of differential DNA methylation in normal and transformed human cells**. *Nature Genetics*. Nature Publishing Group; 37: 853. doi:10.1038/ng1598

[131] Serre D, Lee BH, Ting AH (2010): **MBD-isolated Genome Sequencing provides a high-throughput and comprehensive survey of DNA methylation in the human genome**. *Nucleic Acids Research*. Oxford University Press; 38: 391–399. doi:10.1093/nar/gkp992

[132] Frommer M, McDonald LE, Millar DS, Collis CM, Watt F, Grigg GW, Molloy PL, Paul CL (1992): **A genomic sequencing protocol that yields a positive display of 5-methylcytosine residues in individual DNA strands.** *Proceedings of the National Academy of Sciences of the United States of America*. National Academy of Sciences; 89: 1827–1831. doi:10.1073/pnas.89.5.1827

[133] Huang Y, Pastor WA, Shen Y, Tahiliani M, Liu DR, Rao A (2010): **The behaviour of 5-hydroxymethylcytosine in bisulfite sequencing**. *PLoS One*. 5: e8888. doi:10.1371/journal.pone.0008888

[134] Cock PJA, Fields CJ, Goto N, Heuer ML, Rice PM (2010): **The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants**. *Nucleic Acids Research*. Oxford University Press; 38: 1767–1771. doi:10.1093/nar/gkp1137

[135] Ewing B, Green P (1998): **Base-calling of automated sequencer traces using PHRED. II. Error probabilities**. *Genome research*. 8: 186–194. Available: https://www.ncbi.nlm.nih.gov/pubmed/9521922

[136] Andrews S (2018.): **Babraham Bioinformatics - FastQC A Quality Control tool for High Throughput Sequence Data** [Internet]. Available: http://www.bioinformatics.babraham.ac.uk/projects/fastqc

[137] Bolger AM, Lohse M, Usadel B (2014): **Trimmomatic: a flexible trimmer for Illumina sequence data**. *Bioinformatics*. Oxford University Press; 30: 2114–2120. doi:10.1093/bioinformatics/btu170

[138] Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL (2009): **BLAST+: Architecture and applications.** *BMC bioinformatics*. 10: 421. doi:10.1186/1471-2105-10-421

[139] Flicek P, Birney E (2009): **Sense from sequence reads: methods for alignment and assembly**. *Nature Methods*. Nature Publishing Group; 6: S6. doi:10.1038/nmeth.1376

[140] Kent WJ (2002): **BLAT – The BLAST-Like Alignment Tool**. *Genome Research*. Cold Spring Harbor Lab; 12: 656–664. doi:10.1101/gr.229202

[141] Morgulis A, Coulouris G, Raytselis Y, Madden TL, Agarwala R, Schäffer AA (2008): **Database indexing for production MegaBLAST searches**. *Bioinformatics*. Oxford University Press; 24: 1757–1764. doi:10.1093/bioinformatics/btn322

[142] Buchfink B, Xie C, Huson DH (2014): **Fast and sensitive protein alignment using DIAMOND**. *Nature Methods*. Nature Publishing Group; 12: 59. doi:10.1038/nmeth.3176

[143] Oehmen C, Nieplocha J (2006): **ScalaBLAST: A Scalable Implementation of BLAST for High-Performance Data-Intensive Bioinformatics Analysis**. *IEEE Transactions on Parallel and Distributed Systems*. IEEE; 17: 740–749. doi:10.1109/TPDS.2006.112

[144] Vouzis PD, Sahinidis NV (2011): **GPU-BLAST: using graphics processors to accelerate protein sequence alignment**. *Bioinformatics*. Oxford University Press; 27: 182–188. doi:10.1093/bioinformatics/btq644

[145] Li R, Li Y, Kristiansen K, Wang J (2008): **SOAP: short oligonucleotide alignment program**. *Bioinformatics*. Oxford University Press; 24: 713–714. doi:10.1093/bioinformatics/btn025

[146] Li H, Ruan J, Durbin R (2008): **Mapping short DNA sequencing reads and calling variants using mapping quality scores**. *Genome Research*. Cold Spring Harbor Lab; 18: 1851–1858. doi:10.1101/gr.078212.108

[147] Langmead B, Trapnell C, Pop M, Salzberg SL (2009): **Ultrafast and memory-efficient alignment of short DNA sequences to the human genome**. *Genome Biology*. BioMed Central; 10: R25. doi:10.1186/gb-2009-10-3-r25

[148] Li H, Durbin R (2009): **Fast and accurate short read alignment with Burrows–Wheeler transform**. *Bioinformatics*. Oxford University Press; 25: 1754–1760. doi:10.1093/bioinformatics/btp324

[149] Li R, Yu C, Li Y, Lam T-W, Yiu S-M, Kristiansen K, Wang J (2009): **SOAP2: an improved ultrafast tool for short read alignment**. *Bioinformatics*. Oxford University Press; 25: 1966–1967. doi:10.1093/bioinformatics/btp336

[150] Burrows M, Wheeler DJ (1994): **A block-sorting lossless data compression algorithm**. *SRC Research Report*. Digital Systems Research Center;

[151] Hoffmann S, Otto C, Kurtz S, Sharma CM, Khaitovich P, Vogel J, Stadler PF, Hackermüller J (2009): **Fast mapping of short sequences with mismatches, insertions and deletions using index structures**. *PLoS Computational Biology*. 5: e1000502. doi:10.1371/journal.pcbi.1000502

[152] Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R (2009): **The Sequence Alignment/Map format and SAMtools**. *Bioinformatics*. Oxford University Press; 25: 2078–2079. doi:10.1093/bioinformatics/btp352

[153] Hoffmann S, Otto C, Doose G, Tanzer A, Langenberger D, Christ S, Kunz M, Holdt LM, Teupser D, Hackermüller J, Stadler PF (2014): **A multi-split mapping algorithm for circular RNA, splicing, trans-splicing and fusion detection**. *Genome biology*. 15: R34. doi:10.1186/gb-2014-15-2-r34

[154] Langmead B, Salzberg SL (2012): **Fast gapped-read alignment with Bowtie 2**. *Nature Methods*. Nature Publishing Group; 9: 357. doi:10.1038/nmeth.1923

[155] Trapnell C, Pachter L, Salzberg SL (2009): **TopHat: discovering splice junctions with RNA-Seq**. *Bioinformatics*. Oxford University Press; 25: 1105–1111. doi:10.1093/bioinformatics/btp120

[156] Ghosh S, Chan C-KK (2016): **Analysis of RNA-Seq Data Using TopHat and Cufflinks**. *Methods of Molecular Biology*. Springer; 1374: 339–61. doi:10.1007/978-1-4939-3167-5_18

[157] Wang K, Singh D, Zeng Z, Coleman SJ, Huang Y, Savich GL, He X, Mieczkowski P, Grimm SA, Perou CM, MacLeod JN, Chiang DY, Prins JF, Liu J (2010): **MapSplice: accurate mapping of RNA-seq reads for splice junction discovery**. *Nucleic Acids Research*. Silverchair Information Systems; 38: e178. doi:10.1093/nar/gkq622

[158] Au KF, Jiang H, Lin L, Xing Y, Wong WH (2010): **Detection of splice junctions from paired-end RNA-seq data by SpliceMap**. *Nucleic Acids Research*. Silverchair Information Systems; 38: 4570–4578. doi:10.1093/nar/gkq211

[159] Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR (2013): **STAR: ultrafast universal RNA-seq aligner**. *Bioinformatics*. Silverchair Information Systems; 29: 15–21. doi:10.1093/bioinformatics/bts635

[160] Otto C, Stadler PF, Hoffmann S (2014): **Lacking alignments? The next-generation sequencing mapper segemehl revisited**. *Bioinformatics*. Oxford University Press; 30: 1837–1843. doi:10.1093/bioinformatics/btu146

[161] Kolkman JA, Stemmer WPC (2001): **Directed evolution of proteins by exon shuffling**. *Nature Biotechnology*. Nature Publishing Group; 19: 423. doi:10.1038/88084

[162] Danan M, Schwartz S, Edelheit S, Sorek R (2011): **Transcriptome-wide discovery of circular RNAs in archaea**. *Nucleic Acids Research*. Oxford University Press; 40: 3131–3142. doi:10.1093/nar/gkr1009

[163] Hansen TB, Venø MT, Damgaard CK, Kjems J (2016): **Comparison of circular RNA prediction tools**. *Nucleic Acids Research*. Oxford Univ Press; 44: e58–e58. doi:10.1093/nar/gkv1458

[164] Zeng X, Lin W, Guo M, Zou Q (2017): **A comprehensive overview and evaluation of circular RNA detection tools**. *PLoS Computational Biology*. Public Library of Science; 13: e1005420. doi:10.1371/journal.pcbi.1005420

[165] Zhang X-O, Wang H-B, Zhang Y, Lu X, Chen L-L, Yang L (2014): **Complementary Sequence-Mediated Exon Circularization**. *Cell*. Cell Press; 159: 134–147. doi:10.1016/j.cell.2014.09.001

[166] Gao Y, Wang J, Zhao F (2015): **CIRI: An efficient and unbiased algorithm for de novo circular RNA identification**. *Genome Biology*. BioMed Central; 16: 4. doi:10.1186/s13059-014-0571-3

[167] Gao Y, Zhang J, Zhao F (2018): **Circular RNA identification based on multiple seed matching**. *Briefings in Bioinformatics*. Oxford University Press; 19: 803–810. doi:10.1093/bib/bbx014

[168] Otto C, Stadler PF, Hoffmann S (2012): **Fast and sensitive mapping of bisulfite-treated sequencing data**. *Bioinformatics*. Oxford University Press; 28: 1698–1704. doi:10.1093/bioinformatics/bts254

[169] Krueger F, Andrews SR (2011): **Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications**. *Bioinformatics (Oxford, England)*. Silverchair Information Systems; 27: 1571–1572. doi:10.1093/bioinformatics/btr167

[170] Zhao S, Zhang Y, Gordon W, Quan J, Xi H, Du S, Schack D von, Zhang B (2015): **Comparison of stranded and non-stranded RNA-seq transcriptome profiling and investigation of gene overlap**. *BMC Genomics*. BioMed Central; 16: 675. doi:10.1186/s12864-015-1876-7

[171] Liao Y, Smyth GK, Shi W (2014): **featureCounts: an efficient general purpose program for assigning sequence reads to genomic features**. *Bioinformatics*. Oxford University Press; 30: 923–930. doi:10.1093/bioinformatics/btt656

[172] Bray NL, Pimentel H, Melsted P, Pachter L (2016): **Near-optimal probabilistic RNA-seq quantification**. *Nature Biotechnology*. Nature Publishing Group; 34: 525. doi:10.1038/nbt.3519

[173] Love MI, Huber W, Anders S (2014): **Moderated estimation of fold change and dispersion for RNA-seq data with deseq2**. *Genome Biology*. BioMed Central; 15: 550. doi:10.1186/s13059-014-0550-8

[174] Benjamini Y, Hochberg Y (1995): **Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing**. *Journal of the Royal Statistical Society Series B (Methodological)*. Royal Statistical Society, Wiley; 57: 289–300. doi:10.2307/2346101

[175] Mandin P, Toledo-Arana A, D'Hérouel AF, Repoila F (2013): **RNA-mediated Control of Bacterial Gene Expression: Role of Regulatory non-Coding RNAs**. *American Cancer Society*. American Cancer Society; 1–36. doi:10.1002/3527600906.mcb.201200016

[176] Freyhult EK, Bollback JP, Gardner PP (2007): **Exploring genomic dark matter: A critical assessment of the performance of homology search methods on noncoding RNA**. *Genome Res*. Cold Spring Harbor Lab; 17: 117–125. doi:10.1101/gr.5890907

[177] Wilbur WJ, Lipman DJ (1983): **Rapid similarity searches of nucleic acid and protein data banks**. *Proceedings of the National Academy of Sciences*. National Academy of Sciences; 80: 726–730. doi:10.1073/pnas.80.3.726

[178] Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, McWilliam H, Valentin F, Wallace IM, Wilm A, Lopez R, Thompson JD, Gibson TJ, Higgins DG (2007): **Clustal W and Clustal X version 2.0**. *Bioinformatics*. Oxford University Press; 23: 2947–2948. doi:10.1093/bioinformatics/btm404

[179] Sievers F, Wilm A, Dineen D, Gibson TJ, Karplus K, Li W, Lopez R, McWilliam H, Remmert M, Söding J, others (2011): **Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega**. *Molecular Systems Biology*. EMBO Press; 7: 539. doi:10.1038/msb.2011.75

[180] Nawrocki EP, Eddy SR (2013): **Infernal 1.1: 100-fold faster RNA homology searches**. *Bioinformatics*. Oxford University Press; 29: 2933–2935. doi:10.1093/bioinformatics/btt509

[181] Will S, Joshi T, Hofacker IL, Stadler PF, Backofen R (2012): **LocARNA-p: Accurate boundary prediction and improved detection of structural RNAs**. *RNA*. Cold Spring Harbor Lab; 18: 900–914. doi:10.1261/rna.029041.111

[182] Nawrocki EP, Burge SW, Bateman A, Daub J, Eberhardt RY, Eddy SR, Floden EW, Gardner PP, Jones TA, Tate J, others (2014): **Rfam 12.0: Updates to the RNA families database**. *Nucleic Acids Research*. Oxford University Press; 43: D130–D137. doi:10.1093/nar/gku1063

[183] Lindgreen S, Umu SU, Lai AS-W, Eldai H, Liu W, McGimpsey S, Wheeler NE, Biggs PJ, Thomson NR, Barquist L, Poole AM, Gardner PP (2014): **Robust Identification of Noncoding RNA from Transcriptomes Requires Phylogenetically-Informed Sampling**. *PLoS Comput Biol*. Public Library of Science; 10: e1003907. doi:10.1371/journal.pcbi.1003907

[184] Lorenz R, Bernhart SH, Zu Siederdissen CH, Tafer H, Flamm C, Stadler PF, Hofacker IL (2011): **ViennaRNA package 2.0**. *Algorithms for Molecular Biology*. BioMed Central; 6: 26.

[185] Wolfinger MT, Svrcek-Seiler WA, Flamm C, Hofacker IL, Stadler PF (2004): **Efficient computation of RNA folding dynamics**. *Journal of Physics A: Mathematical and General*. IOP Publishing; 37: 4731. doi:10.1088/0305-4470/37/17/005

[186] Zuker M (2003): **Mfold web server for nucleic acid folding and hybridization prediction**. *Nucleic acids research*. Oxford University Press; 31: 3406–3415. doi:10.1093/nar/gkg595

[187] Rehmsmeier M, Steffen P, Höchsmann M, Giegerich R (2004): **Fast and effective prediction of microRNA/target duplexes**. *RNA*. Cold Spring Harbor Lab; 10: 1507–1517. doi:10.1261/rna.5248604

[188] Wright PR, Georg J, Mann M, Sorescu DA, Richter AS, Lott S, Kleinkauf R, Hess WR, Backofen R (2014): **CopraRNA and IntaRNA: predicting small RNA targets, networks and interaction domains**. *Nucleic Acids Res*. Oxford University Press; 42: W119–W123. doi:10.1093/nar/gku359

[189] Mann M, Wright PR, Backofen R (2017): **IntaRNA 2.0: enhanced and customizable prediction of RNA–RNA interactions**. *Nucleic Acids Res*. Oxford University Press; 45: W435–W439. doi:10.1093/nar/gkx279

[190] Santangelo TJ, Artsimovitch I (2011): **Termination and antitermination: RNA polymerase runs a stop sign**. *Nat Rev Microbiol*. Nature Publishing Group; 9: 319. doi:10.1038/nrmicro2560

[191] Stamatakis A (2014): **RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies**. *Bioinformatics*. Oxford University Press; 30: 1312–1313. doi:10.1093/bioinformatics/btu033

[192] Henikoff S, Henikoff JG (1992): **Amino acid substitution matrices from protein blocks**. *Proceedings of the National Academy of Sciences*. National Academy of Sciences; 89: 10915–10919. doi:10.1073/pnas.89.22.10915

[193] Le SQ, Gascuel O (2008): **An Improved General Amino Acid Replacement Matrix**. *Molecular Biology and Evolution*. Oxford University Press; 25: 1307–1320. doi:10.1093/molbev/msn067

[194] Darriba D, Taboada GL, Doallo R, Posada D (2011): **ProtTest 3: fast selection of best-fit models of protein evolution**. *Bioinformatics*. Oxford University Press; 27: 1164–1165. doi:10.1093/bioinformatics/btr088

[195] Efron B (1979): **Bootstrap Methods: Another Look at the Jackknife**. *Annals of Statistics*. Institute of Mathematical Statistics; 7: 1–26. doi:10.1214/aos/1176344552

[196] Leinonen R, Sugawara H, Shumway M, Collaboration INSD (2010): **The Sequence Read Archive**. *Nucleic Acids Research*. Oxford University Press; 39: D19–D21. doi:10.1093/nar/gkq1019

[197] Gößringer M, Lechner M, Brillante N, Weber C, Rossmanith W, Hartmann RK (2017): **Protein-only RNase P function in *Escherichia coli*: Viability, processing defects and differences between PRORP isoenzymes**. *Nucleic Acids Research*. doi:10.1093/nar/gkx405

[198] Lechner M, Rossmanith W, Hartmann RK, Thölken C, Gutmann B, Giegé P, Gobert A (2015): **Distribution of ribonucleoprotein and protein-only RNase P in eukarya**. *Molecular biology and evolution*. Oxford University Press; 32: 3186–3193. doi:10.1093/molbev/msv187

[199] Yusuf D, Marz M, Stadler PF, Hofacker IL (2010): **Bcheck: A wrapper tool for detecting RNase P RNA genes**. *BMC genomics*. BioMed Central; 11: 432. doi:10.1186/1471-2164-11-432

[200] Edgar RC (2004): **MUSCLE: a multiple sequence alignment method with reduced time and space complexity**. *BMC Bioinformatics*. BioMed Central; 5: 113. doi:10.1186/1471-2105-5-113

[201] Huson DH, Bryant D (2005): **Application of phylogenetic networks in evolutionary studies**. *Molecular biology and evolution*. Oxford University Press; 23: 254–267. doi:10.1093/molbev/msj030

[202] Crooks GE, Hon G, Chandonia J-M, Brenner SE (2004): **WebLogo: A Sequence Logo Generator**. *Genome Research*. Cold Spring Harbor Lab; 14: 1188–1190. doi:10.1101/gr.849004

[203] Zdobnov EM, Tegenfeldt F, Kuznetsov D, Waterhouse RM, Simão FA, Ioannidis P, Seppey M, Loetscher A, Kriventseva EV (2016): **OrthoDB v9. 1: Cataloging evolutionary and functional annotations for animal, fungal, plant, archaeal, bacterial and viral orthologs**. *Nucleic acids research*. Oxford University Press; 45: D744–D749.

[204] Mann HB, Whitney DR (1947): **On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other**. *Annals of Mathematical Statistics*. Institute of Mathematical Statistics; 18: 50–60. doi:10.1214/aoms/1177730491

[205] Kozomara A, Griffiths-Jones S (2014): **MiRBase: Annotating high confidence microRNAs using deep sequencing data**. *Nucleic Acids Research*. 42: D68–D73. doi:10.1093/nar/gkt1181

[206] Enright AJ, John B, Gaul U, Tuschl T, Sander C, Marks DS (2003): **MicroRNA targets in drosophila**. *Genome biology*. BioMed Central; 5: R1.

[207] Nicol JW, Helt GA, Blanchard SG Jr., Raja A, Loraine AE (2009): **The Integrated Genome Browser: free software for distribution and exploration of genome-scale datasets**. *Bioinformatics*. Oxford University Press; 25: 2730–2731. doi:10.1093/bioinformatics/btp472

[208] Robinson JT, Thorvaldsdóttir H, Winckler W, Guttman M, Lander ES, Getz G, Mesirov JP (2011): **Integrative genomics viewer**. *Nature Biotechnology*. Nature Publishing Group; 29: 24. doi:10.1038/nbt.1754

[209] Wolstenholme DR, Macfarlane JL, Okimoto R, Clary DO, Wahleithner JA (1987): **Bizarre tRNAs inferred from DNA sequences of mitochondrial genomes of nematode worms**. *Proceedings of the National Academy of Sciences*. National Academy of Sciences; 84: 1324–1328. doi:10.1073/pnas.84.5.1324

[210] Helm M, Brulé H, Friede D, Giegé R, Pütz D, Florentz C (2000): **Search for characteristic structural features of mammalian mitochondrial tRNAs.** *RNA (New York, NY)*. 6: 1356–1379. doi:10.1017/S1355838200001047

[211] Jarrous N, Reiner R (2007): **Human RNase P: a tRNA-processing enzyme and transcription factor**. *Nucleic Acids Research*. Oxford University Press; 35: 3519–3524. doi:10.1093/nar/gkm071

[212] Gilbert W (1986): **Origin of life: The RNA world**. *Nature*. Nature Publishing Group; 319: 618. doi:10.1038/319618a0

[213] Altman S (2013): **The RNA – Protein World**. *RNA*. Cold Spring Harbor Laboratory Press; 19: 589–590. doi:10.1261/rna.038687.113

[214] Rybak-Wolf A, Stottmeister C, Glažar P, Jens M, Pino N, Giusti S, Hanan M, Behm M, Bartok O, Ashwal-Fluss R, Herzog M, Schreyer L, Papavasileiou P, Ivanov A, Öhman M, Refojo D, Kadener S, Rajewsky N (2015): **Circular RNAs in the mammalian brain are highly abundant, conserved, and dynamically expressed.** *Molecular cell*. 58: 870–885. doi:10.1016/j.molcel.2015.03.027

[215] Liu F, Li W, Li Z, Zhang S, Chen S, Su S (2011): **High-abundance mRNAs in *Apis mellifera*: Comparison between nurses and foragers**. *Journal of Insect Physiology*. Pergamon; 57: 274–279. doi:10.1016/j.jinsphys.2010.11.015

[216] Gandini MA, Felix R (2012): **Functional interactions between voltage-gated Ca 2+ channels and Rab3-interacting molecules (RIMs): New insights into stimulus–secretion coupling**. *Biochimica et Biophysica Acta (BBA)-Biomembranes*. Elsevier; 1818: 551–558. doi:10.1016/j.bbamem.2011.12.011

[217] Garner CC, Kindler S, Gundelfinger ED (2000): **Molecular determinants of presynaptic active zones**. *Current opinion in neurobiology*. Elsevier; 10: 321–327. doi:10.1016/S0959-4388(00)00093-3

[218] Folkers E, Waddell S, Quinn WG (2006): **The Drosophila radish gene encodes a protein required for anesthesia-resistant memory.** *Proceedings of the National Academy of Sciences of the United States of America*. 103: 17496–17500. doi:10.1073/pnas.0608377103

[219] Mery F, Kawecki TJ (2005): **A cost of long-term memory in *Drosophila*.** *Science (New York, NY)*. 308: 1148. doi:10.1126/science.1111331

[220] Tully T, Preat T, Boynton SC, Del Vecchio M (1994): **Genetic dissection of consolidated memory in *Drosophila*.** *Cell*. 79: 35–47.

[221] Wiegmann BM, Trautwein MD, Kim J-W, Cassel BK, Bertone MA, Winterton SL, Yeates DK (2009): **Single-copy nuclear genes resolve the phylogeny of the holometabolous insects.** *BMC biology*. 7: 34. doi:10.1186/1741-7007-7-34

[222] Naeger NL, Van Nest BN, Johnson JN, Boyd SD, Southey BR, Rodriguez-Zas SL, Moore D, Robinson GE (2011): **Neurogenomic signatures of spatiotemporal memories in time-trained forager honey bees.** *The Journal of experimental biology*. 214: 979–987. doi:10.1242/jeb.053421

[223] Weaver DB, Anzola JM, Evans JD, Reid JG, Reese JT, Childs KL, Zdobnov EM, Samanta MP, Miller J, Elsik CG (2007): **Computational and transcriptional evidence for microRNAs in the honey bee genome.** *Genome biology*. 8: R97. doi:10.1186/gb-2007-8-6-r97

[224] Behura SK, Whitfield CW (2010): **Correlated expression patterns of microRNA genes with age-dependent behavioural changes in honeybee**. *Insect Molecular Biology*. 19: 431–439. doi:10.1111/j.1365-2583.2010.01010.x

[225] Chen H, Li Y, Chen K, Yao Q, Li G, Wang L (2010): **Comparative proteomic analysis of *Bombyx mori* hemolymph and fat body after calorie restriction.** *Acta biochimica Polonica*. 57: 505–511.

[226] Liu F, Peng W, Li Z, Li W, Li L, Pan J, Zhang S, Miao Y, Chen S, Su S (2012): **Next-generation small RNA sequencing for microRNAs profiling in *Apis mellifera*: Comparison between nurses and foragers.** *Insect molecular biology*. 21: 297–303. doi:10.1111/j.1365-2583.2012.01135.x

[227] Jeck WR, Sorrentino JA, Wang K, Slevin MK, Burd CE, Liu J, Marzluff WF, Sharpless NE (2013): **Circular RNAs are abundant, conserved, and associated with alu repeats.** *RNA (New York, NY)*. 19: 141–157. doi:10.1261/rna.035667.112

[228] Liang D, Wilusz JE (2014): **Short intronic repeat sequences facilitate circular RNA production**. *Genes & development*. Cold Spring Harbor Lab; 28: 2233–2247. doi:10.1101/gad.251926.114

[229] Kramer MC, Liang D, Tatomer DC, Gold B, March ZM, Cherry S, Wilusz JE (2015): **Combinatorial control of Drosophila circular RNA expression by intronic repeats, hnRNPs, and sr proteins**. *Genes & development*. 29: 2168–2182. doi:10.1101/gad.270421.115

[230] Starke S, Jost I, Rossbach O, Schneider T, Schreiner S, Hung L-H, Bindereif A (2015): **Exon circularization requires canonical splice signals**. *Cell reports*. Elsevier; 10: 103–111. doi:10.1016/j.celrep.2014.12.002

[231] Shukla S, Kavak E, Gregory M, Imashimizu M, Shutinoski B, Kashlev M, Oberdoerffer P, Sandberg R, Oberdoerffer S (2011): **CTCF-promoted rna polymerase ii pausing links dna methylation to splicing**. *Nature*. Nature Publishing Group; 479: 74. doi:10.1038/nature10442

[232] Oka M, Rodić N, Graddy J, Chang L-J, Terada N (2006): **CpG sites preferentially methylated by Dnmt3a *in vivo*.** *The Journal of biological chemistry*. 281: 9901–9908. doi:10.1074/jbc.M511100200

[233] Becker N, Kucharski R, Rössler W, Maleszka R (2016): **Age-dependent transcriptional and epigenomic responses to light exposure in the honey bee brain.** *FEBS open biology*. 6: 622–639. doi:10.1002/2211-5463.12084

[234] Cingolani P, Cao X, Khetani RS, Chen C-C, Coon M, Sammak A, Bollig-Fischer A, Land S, Huang Y, Hudson ME, Garfinkel MD, Zhong S, Robinson GE, Ruden DM (2013): **Intronic non-CG DNA hydroxymethylation and alternative mRNA splicing in honey bees.** *BMC genomics*. 14: 666. doi:10.1186/1471-2164-14-666